ICT 320: Task 2

By Emil Gustafsson

LL

# Table of Contents

[Table of Contents 2](#_Toc22892708)

[1. Current Design 3](#_Toc22892709)

[1.1 Entity Relationship Diagram – Current Design 3](#_Toc22892710)

[1.2 Current Normal Form 3](#_Toc22892711)

[1.3 Multiple Values in Samples 3](#_Toc22892712)

[1.4 Redundant fields 3](#_Toc22892713)

[1.5 Splitting Contacts & Morphs table 4](#_Toc22892714)

[2. Proposed Design 4](#_Toc22892715)

[2.1 Entity Relationship Diagram – Proposed Design 4](#_Toc22892716)

[2.2 Primary & Foreign Keys and Indexes 4](#_Toc22892717)

[3. Assumptions & Supplementary Design Requirements 5](#_Toc22892718)

[3.1 Assumptions 5](#_Toc22892719)

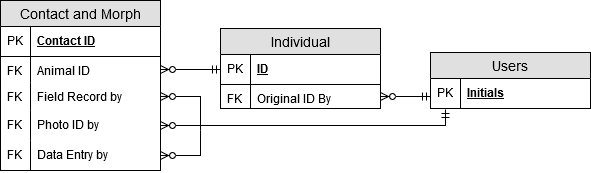
[3.2 Supplementary Design Requirements 5](#_Toc22892720)

[4. Conclusion 5](#_Toc22892721)

LL

# Current Design

## Entity Relationship Diagram – Current Design



Note that current databases field names are in upper case, but here are presented in lowercase for better readability.

## Current Normal Form

The current design does not fulfil the requirements to be in 1NF, because the column “SAMPLE .” in Individuals contains multiple values. And since it does not fulfil 1NF requirements, it cannot be a higher version of Normal Form.

## Multiple Values in Samples

One design flaw is that there are multiple values being stored in the “SAMPLE .”-column from the Individuals table. Because each column should hold the smallest meaningful value and that there should not be multiple values in a single field. The solution is to move these into their own table, called Samples. The alternative would be reintroducing Repeated Fields, like Sample1, Sample2 and Sample3. But this is not good, because problems will occur if for example a fourth sample of an animal is being collected. And each sample should also have a column for, I3SS\_Manta\_Score. That would be at least three more columns, that would most likely contain null values. We should try to organize the data vertically instead of horizontally.

## Redundant fields

We could reduce the number of column and empty fields, by using one ENUM column to combine multiple of the columns in “Contacts & Morphs”.

One sample’s type consists of only one of the following: TAIL, BLOOD, FAECES, SKIN, HEAMATOLOGY, SWAB\_SKIN, SWAB\_ORAL or SWAB\_CLOACAL.

By combining these columns into one ENUM column we will avoid empty fields and unnecessary columns in the database.

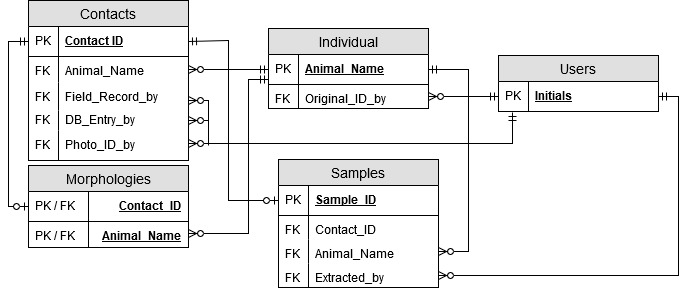
## Splitting Contacts & Morphs table

In the current database Contacts and Morphologies share a table. By this denormalization method, it could save computing power because queries relating to morphologies wouldn’t have to join two tables. But this not great, because a contact may not record morphologies and would therefore result in a lot of NULL values being stored for morphologies, and many unnecessary columns.

In our design we have chosen to split the table. A morphology row is only inserted into the Morphologies table when a contact has Morphology set to TRUE.

# Proposed Design

## Entity Relationship Diagram – Proposed Design



## 2.2 Primary & Foreign Keys and Indexes

The Initials column in Users in the current database are used as the Primary Key in relations to other tables, which we propose to preserve. Even tough the Username column might be better suited, because of its length, it would save a lot of work in converting the data to preserve the current one. And many users seem to fall under categories, like Students or Voluntaries.

We propose changing the Primary Key for Individuals from ID to AnimalName. Firstly, the field is already unique, and therefore a candidate key. We also think the scientists would agree it’s more relatable with a name that’s given instead of memorizing a number for each animal. It would also save computing power in the future, because future queries that wants to display the animal’s names and number of morphologies or contacts wouldn’t have to check multiple tables. Therefore, the column ID is omitted in this design, although an Old\_ID column has replaced its place but no longer functions as a key value. It’s for when converting the old data into our new database.

Morphology’s will be moved to its own table, with two unique Foreign Keys as its Primary Key. Because there is no need for a unique ID, since the table is always related to a contact with one animal.

The Sample columns has been moved to its own table, primarily identified by Sample\_ID. We could instead have used unique foreign keys to identify the sample and its result values, like in the Morphology table. Because one sample only belongs to one contact. But since Samples are identified by numbers or characters, unlike our Morphologies table, it’s the best option.

We have decided to create Indexes for the columns Morphology in Individuals and DNA in Contacts. Because the data sparsity in the columns is low, and the values will be used in WHERE clauses in future queries.

# Assumptions & Supplementary Design Requirements

## Assumptions

* We have assumed that a sample is one of either TAIL, BLOOD, FAECES, SKIN, HEAMATOLOGY, SWAB\_SKIN, SWAB\_ORAL or SWAB\_CLOACAL.
* We have assumed that one morphology in "Number of morphologies” constitutes if the column “MORPHOLOGY?” is True in one Contact. As opposed to one morphology measurement column having value in one Contact constituting one morphology.
* We have assumed that Contact & Morphs rows in the old data not containing reference to an actual animal are irrelevant and therefore omitted.
* We have assumed that max one type of Sample is collected during one contact with the animal.

## Supplementary Design Requirements

* Some ENUM columns in our design has NULL values as a choice, to comply with the old data where some rows are missing any of the specified choices.
* When importing the data from the old database, we recommend using the user Unknown (added in our design) in the rows that is missing foreign keys to the table Users. Otherwise these rows would be invalid and skipped, resulting in loss of otherwise valid data.
* In this database design there is a row called Old\_ID in the Individuals table. This is needed for correctly inserting and correctly relating all the old data into our new system. After running our script, and all the data is inserted and related, this column can be deleted.
* There are a couple of rows in the database that is missing Animal Name or ID in the old data, and in our test conversion we omitted these. But if the scientists deem the rows relevant nonetheless, they could be preserved by creating a row for Unknown animals and use that instead. Much like our user Unknown.
* We have not specified Integer Width when declaring columns, because the function is deprecated and will be removed in a future version of MySQL.

# Conclusion

By using the Animal\_Name as the Primary Key we can avoid unnecessary joins in future queries that want to display the name with, for example, number of morphologies. We could also have denormalized the affected tables by adding a duplicate field of Animal\_Name and keep the ID as Primary Key. But then we need to insert or update data in multiple places, which increases the chances of data anomalies.

And by splitting the Contacts & Morphs table, we have greatly reduced the unused space in the database. Contact’ have six less columns now, and a Morphology is only inserted into the database when there is such data. This means a lot of less written rows, and a more efficient database.

We also moved the Samples into its own table, which brings the same advantages as the splitting of Contacts & Morphs table. We have reduced the amount of data written into our database, since Samples aren’t always collected, we no longer need to write any empty rows or redundant FALSE fields for contacts without any collected sample.

We have also replaced several redundant columns regarding the Sample type with an ENUM column, additionally saving a lot of storage and space. Instead of having seven “False” values and one “True” if a Sample is collected during a contact, or eight “False” values if not, we instead create a Sample-row only when a Sample acutely is collected. And in that row, only one row is related to the type of sample, instead of eight.

We have declared indexes for the Morphology-column in Individuals, and the DNA-column in the Contacts-table. This is because those fields will be appearing in where clauses in future queries as described in the use case description, and that their data sparsity is low.