

An introduction to database migration

Summary	2
Introduction	2
Diagnose data quality before migration	4
Migration steps	7
Before migration: planning	7
Data profiling	7
Data mapping	8
Allocating time and resources	10
During migration	11
Data extraction	11
Data transformation	11
Validation and testing	16
Data loading	17
After the migration	17

Summary

This guide is intended to provide some guidance and tips for human rights practitioners who are carrying out database migration processes.

Data migration could be tough if we do not take the time to organize and consciously evaluate what, why and how to migrate in advance. Destroying or losing relevant information during the process, or migrating dirty or inconsistent data already stored are common risks related to data migration. In order to successfully migrate your data from an existing database to a target one, following, we offer some advice that we hope it will enable you to properly diagnose your database, plan and implement your migration project.

Introduction

Database migration is the process of moving data from a source database to a target database.

Diagram 1: Data migration process



Moving data from one place to another is a recurrent necessity that pursues the improvement of data management performance. More specifically, reasons for data migration could be as diverse as: improving accessibility, security or governance; migrating to a new application; replacing equipment to prevent data loss; updating formats to avoid the obsolescence of technologies, etc.

These purposes are intimately related with the characteristics of the source system and the target system. Regarding the infrastructure, it is common to move from an on-premise database to a cloud database to improve data accessibility or management efficiency. If the database is already online, organizations might want to migrate to new software solutions looking for new features.

Since a database is a collection of structured data, migration does not only mean moving data, but loading it into a new structure after a transformation process that encompasses data cleaning and data mapping (matching fields from different data models).

This transformation can take place in origin (preparation) or after the extraction (transformation). Once the process is finished, the data should be fully located in the target database keeping its previous quality; in other words, the features that make it usable. The source database is often discarded after the migration process, when the data migrated is fully verified.

Undertaking this process manually can be effort consuming and tedious. Furthermore, it could lead to mistakes such as record duplication or data entry inconsistencies, mostly when working with large datasets collected over years. To the extent possible, the extraction, transformation and loading operations should be automated to ensure consistency. Sometimes, automated data migration requires advanced technological skills, but depending on the magnitude of the project there are basic transformation strategies to reformat the data and have it ready for the load.

The steps and tasks shown below are valid to migrate large and small amounts of data, but big migration projects might require additional planning and control measures. If the data comes from different sources or the process is carried out sequentially, it is very important to keep track of every step to avoid duplicating work or losing data.

It is also important to remark that the complexity of the project is not only determined by the amount of data to migrate. As we will see later, the data quality and the differences between the source and the target databases can add new levels of complexity to the process.

Integrating data from different sources is often considered as independent to data migration, but both processes share the extraction, transformation and loading procedure, so they can be undertaken together. The more different the datasets to combine are, more

transformation actions would be required to successfully match them within the new system.

Diagnose data quality before migration

When it comes to migrating data, it is important to diagnose whether your dataset is ready for such a task. The concept of data quality refers to the set of properties that determine data fitness for use. Although data quality is a bigger independent issue than migration, this instance represents a great opportunity to check and if necessary enhance your dataset for the new target database.

Quality encompasses several characteristics, directly linked with the data purposes, so they can be prioritized differently depending on each organization's perspective. Despite this, these are some general dimensions that should be considered¹:

Diagram 2: Data quality dimensions



¹ Taken from "The Six primary dimensions for data quality assessment". Data Management Association - DAMA. United Kingdom, 2013.

Accuracy. Does the data describe the reality as it is? Accurate data describes the object represented in the database with the right values, both in form and format.

A typical mistake regarding format accuracy is recording dates with wrong format. If the database format is MM/DD/YYYY and someone registers dates with the DD/MM/YYYY format, the data stored in the database won't reflect reality correctly.

Uniqueness. Is the data recorded with no redundancies? Avoid unnecessary duplication of information as much as possible. As an efficiency principle, each piece of data should be recorded once, which means that the number of objects in the real world matches the number of entities represented in the database.

If the database registers victims of violent events by capturing biographical and socioeconomic characteristics, each victim should be recorded once. If the same victim suffers two different violent events and gets registered twice, there will be a duplicity. To ensure uniqueness, each entity—in this example, the victim—is identified with a primary key that can be related with various violent events.

Timeliness. Is the data available when it is required? How long does it take from when an event is noticed until its data is properly registered? How much time passes from data collection to database recording?

Completeness. Are all the required data items properly recorded? How many blank values are there in the database? Blank values are empty slots in the database, which means that there is no value for a given property. A high percentage of blanks in the same category might be a signal of a broader issue, as misunderstanding of its meaning or need of reformulation.

If the database is supposed to help to provide telephonic assistance to victims but the telephone of a victim is not recorded, the database will fail in its mission due to lack of data.

Consistency. Are there any contradictions between data? Does the data match across different data stores?

If the database registers the victim's date of birth and, also, the victim's age within a range of ages, there should be no contradictions between both fields.

Validity. Is the data valid according to the indexing rules? Databases often include syntax rules to set the data format, typology and range. Some common data types are text,

numeric, date, time, etc. Formats, for its part, determine how values are displayed; for instance, how to record dates or names of persons.

To reduce errors, it is common to ban letters in phone numbers, define the length of the ID field, set maximum or minimum values, etc. If the database is designed with these constraints, it is not possible to enter data if it does not comply with rules. This is important to consider when analyzing the target database.

Understanding these dimensions is relevant to plan the migration because the process could magnify data quality flaws. The inadvertent presence of duplicate records could lead to replicating this error or deleting valuable information. The syntax rules of both, the legacy database and the new database, are also essential for the migration, as data could be rejected in the new system if it does not meet format, typology or range rules.

Additionally, assessing data quality is the only way to correctly estimate costs and time, and allocate resources for the process. However, foremost is the fact that migration is supposed to maintain data quality, so it is essential to identify quality baselines and goals.

Frequently, data quality is also related to external features, such as *accessibility*, *trustworthiness* and *security*, which also constitute common reasons for database migration.

- **Accessibility** refers to the availability of different pieces of data for those who need it when they need it. As mentioned, this is a recurring goal when migrating offline databases to online platforms.
- **Trustworthiness** depends on the whole data management flow, but regarding data migration is achieved by carefully documenting every transformation in the dataset and who is accountable for each action.
- **Security** is a major reason for database migration in the human rights field. It is common to keep high security standards in the source database and plan to keep or increase them in the target database. However, security is sometimes forgotten in the intermediate steps, when we generate backups. These intermediate files, as well as the retiring systems, should be wiped out once the content is verified at the target database. As always, the best way to reduce security threats is addressing these challenges at the planning stage.

Migration steps

1. Before migration: planning

Planning is the best way to ensure the migration's viability and that it will lead to the expected results in terms of data quality. Once the data quality objectives are clear, it is time to analyze the dataset and the source database. This includes answering questions such as:

- What is the volume of data to migrate? Is it a full migration or it is a partial migration? What are the criteria to select the data to be migrated? Does the data come from one or various sources? Are all the potential data sources identified?
- What is the situation of data in terms of quality (accuracy, uniqueness, timeliness, completeness, consistency and validity)? This analysis will allow us to identify quality gaps, but it is still required to assess if the gaps are solvable during the migration, and at what cost. Going back to the victim's database example, the analysis could reveal that the *phone* field is empty for one third of victims, but the migration won't be able to solve this problem. However, if duplicate victims are identified, it will be possible to reduce redundancy by eliminating one record or merging both.

Data profiling

The previous questions are part of data profiling, a set of techniques aiming to provide insights about the quality of the dataset. Data profiling allows us to quickly have a general overview of the dataset which often includes statistics, summaries of data types and patterns, blank values, etc. These are some of the most common statistics calculated for each column of the dataset (each column represents an attribute or property describing the entity):

- o Number of unique values and distinct values
- o Maximum and minimum values
- o Sum of values
- o Number of blank values
- o Mean, median, mode and range

You can learn more about this on [this course offered by Advocacy Assembly and School of Data](#). It is possible to get simple descriptive statistics of the dataset with most spreadsheet softwares:

- [Descriptive Statistics with LibreOffice Calc](#)
- [Data profiling tools included with Microsoft Excel](#)

There are more questions that you will need to answer by carefully examining your source database, but also asking the team members who use the data regularly. The migration process is a great opportunity to better understand the source database and overcome its limitations with the new system. These are some additional questions you should answer:

- Is it possible to label each piece of data with its typology and format? Is there an identifiable format for dates, ID codes, names, etc.? Is it consistent along the dataset?
- What is the data structure? Is it possible to identify the type of objects represented in the dataset and the relations between them?

Data mapping

The target database needs to accommodate the data once it is cleaned and transformed. Data mapping determines how fields from the source and the target database will match, establishing a relationship between two or more data structures. This process includes the structure, the typology and the syntax rules that apply to the data in order to make connections between the source and the target databases.

The example below shows a simple data mapping diagram in which several data stores from the source database are merged into a new container at the target database:

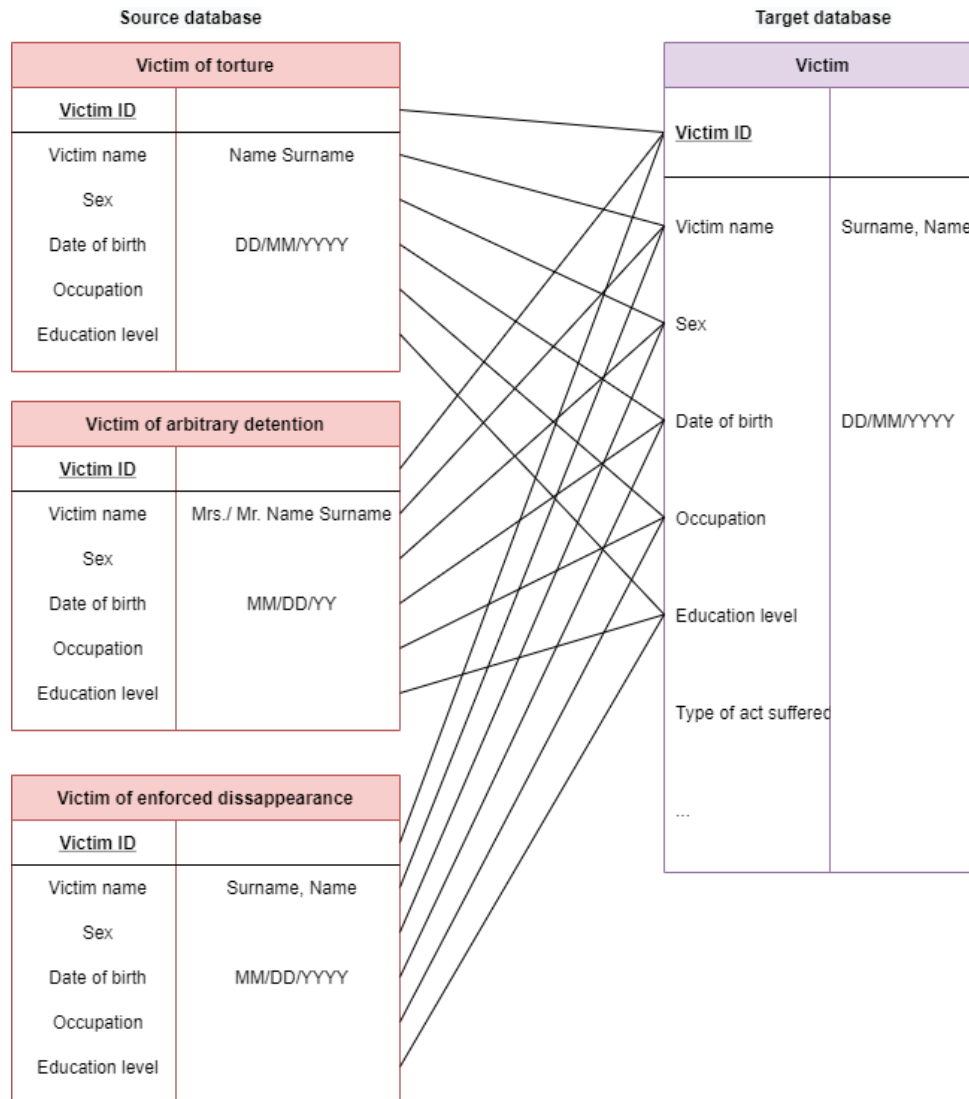


Diagram 3. Basic data mapping model

At the source database, each type of victim is recorded separately according to the type of act suffered (torture, arbitrary detention and enforced disappearance). The target database, on the other hand, is designed to record all the victims in the same bucket.

In this case the matching is very simple, as the different victim typologies share the same fields. After matching fields, we just need to add the field *type of act suffered* (to select between torture, arbitrary detention and enforced disappearance) and we will have the same information in the source and the target database.

However, we still need to define how formats would match. In the source database the fields *victim name* and *date of birth* are formatted differently for each type of victim, so it will be necessary to decide the format at the target database and make the transformations, if applicable.

Similarly, the syntax rules at the target database should not be incompatible with the values at the source database (at least, without being transformed) as this could result in a loss of data. For instance, if the *Victim ID* field in the source database ranges from 5 to 9 digits and in the target database is limited to 7 digits, it is very likely to lose data. In this case, data profiling could help us to anticipate validation problems.

All the questions and techniques mentioned above help to identify the challenges of the migration and, consequently, to allocate resources for it. Data profiling and data mapping are essential to recognize threats and select the best option to tackle them.

Allocating time and resources

On the basis of the analysis results and the project goals, it will be possible to estimate time and costs for the migration project. Although most of the steps and techniques presented in this document are useful for every project, regardless of its level of complexity, complex migrations might require additional human resources or specific tools to succeed.

Furthermore, depending on how migration is designed, it is likely that data won't be available during the process. It is recommendable to notify everyone who uses the data about the process stages and the availability of the data on each one.

Security can be a goal itself for database migration, but it also constitutes a basic issue to consider when planning the project. In order to prevent data loss during manipulation, it is recommended to backup the database at the very beginning of the process. Additionally, the same access permissions that apply to the source database should be kept during the migration, paying special attention if third parties are involved in the process. As mentioned above, intermediate or testing files should be carefully wiped out to avoid data leaks.

In general, an exhaustive planning considers all the relevant factors for the success of the project, which might be different for each organization and each migration purpose. Therefore, spending enough time at this stage is the best investment for the project's success.

2. During migration

Data extraction

Collecting the data you want to migrate from one or more sources is the first step in the migration process itself. To do so, you need to identify all your data sources, as information is usually stored in different systems. For example, you can have a database where you have consolidated human rights violation events in the last decade, but you still have records from the previous years in paper. Or maybe you are partnering with an organization that sends you this data from a remote region where you don't have your own staff, but this data is registered in a spreadsheet. Migration can be a good opportunity to consolidate data, especially if the records are similar but are registered in different formats, as in this example.

If you still have paper records it is very likely that you will need to transcribe them manually on a computer (if the number of records is very large [you can consider acquiring an OCR software](#)). However, if you are working with digital files you should check if your source application (the one where your data is stored) allows you to download the dataset in an interoperable format.

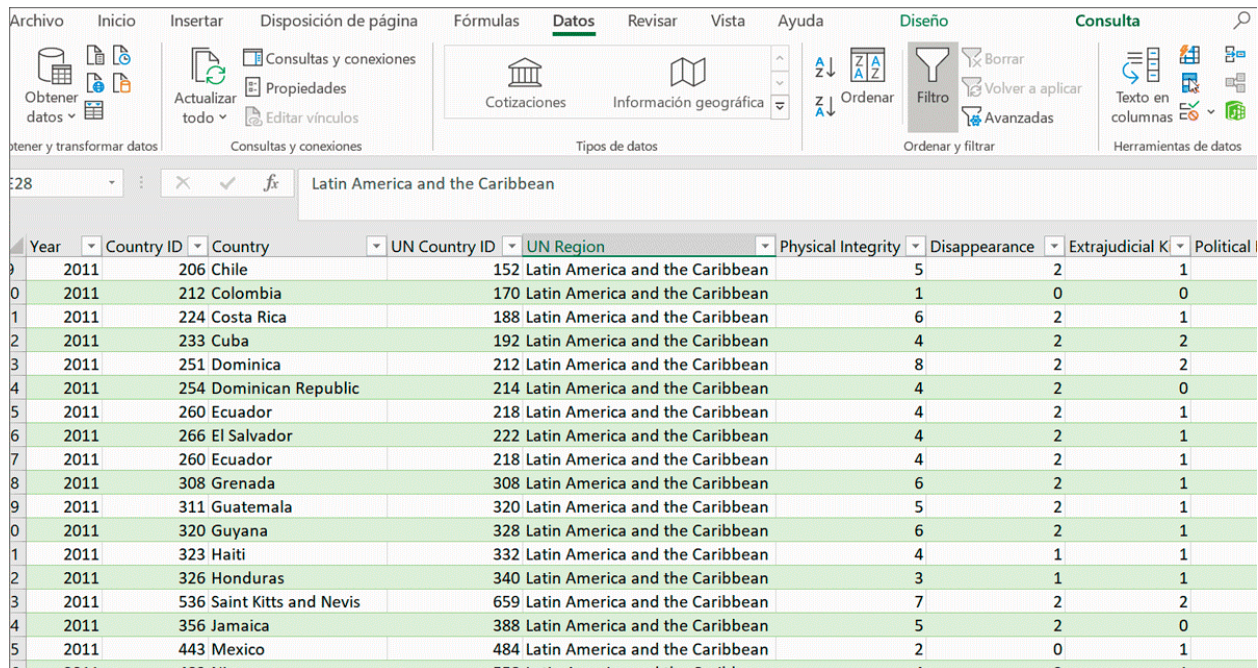
The CSV format is one of the best choices as it saves the values and its structure in a simple plain text machine-readable format. The CSV files contain values which would normally be displayed in columns, but instead are separated by commas. For further information about [how to import and export csv files we recommend you to watch this video](#).

Data transformation

After the extraction, we should have all the data we want to migrate saved in one or more interoperable files, so we can transform them with specialized tools, such as [OpenRefine](#), or simply with a spreadsheet software. The goal of this process is to get your data ready for the load, in the output format and structure.

As data transformation and data cleaning constitute broad and highly specialized fields of work, in this document we will focus on basic steps you can follow to have your data as tidy as possible before the loading (please note that there are several methods that you can use for the same purposes).

Removing duplicate pieces of data:

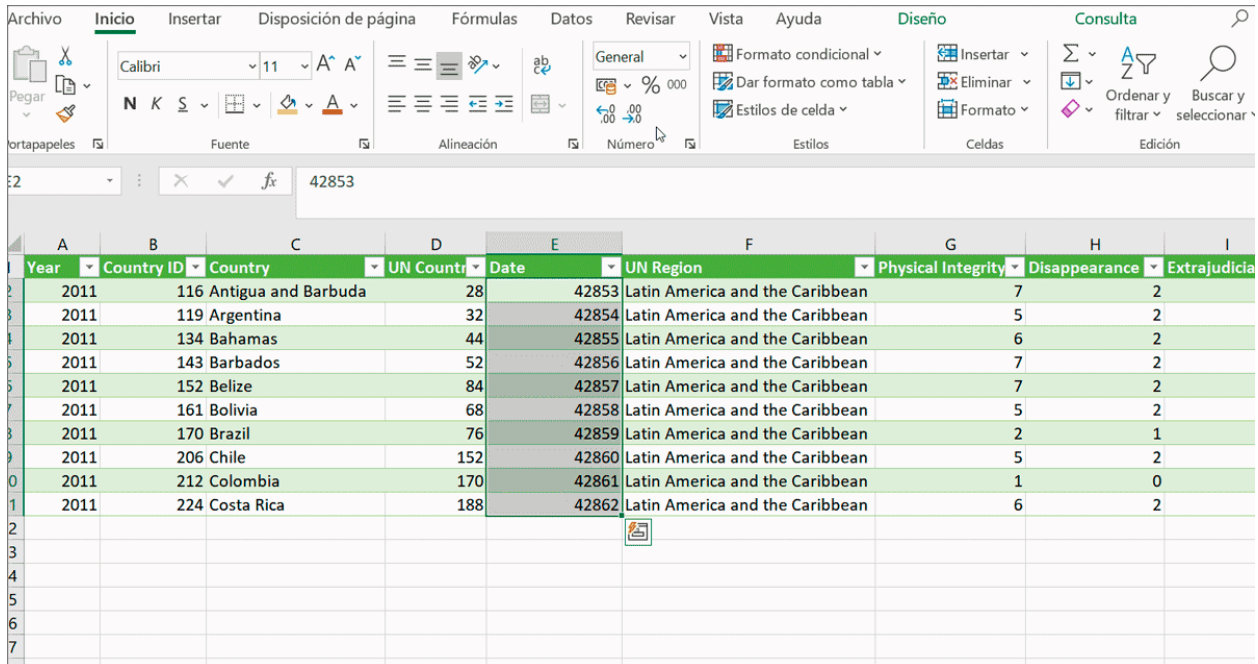


Year	Country ID	Country	UN Country ID	UN Region	Physical Integrity	Disappearance	Extrajudicial Killings	Political Violence
2011	206	Chile	152	Latin America and the Caribbean	5	2	1	
2011	212	Colombia	170	Latin America and the Caribbean	1	0	0	
2011	224	Costa Rica	188	Latin America and the Caribbean	6	2	1	
2011	233	Cuba	192	Latin America and the Caribbean	4	2	2	
2011	251	Dominica	212	Latin America and the Caribbean	8	2	2	
2011	254	Dominican Republic	214	Latin America and the Caribbean	4	2	0	
2011	260	Ecuador	218	Latin America and the Caribbean	4	2	1	
2011	266	El Salvador	222	Latin America and the Caribbean	4	2	1	
2011	260	Ecuador	218	Latin America and the Caribbean	4	2	1	
2011	308	Grenada	308	Latin America and the Caribbean	6	2	1	
2011	311	Guatemala	320	Latin America and the Caribbean	5	2	1	
2011	320	Guyana	328	Latin America and the Caribbean	6	2	1	
2011	323	Haiti	332	Latin America and the Caribbean	4	1	1	
2011	326	Honduras	340	Latin America and the Caribbean	3	1	1	
2011	536	Saint Kitts and Nevis	659	Latin America and the Caribbean	7	2	2	
2011	356	Jamaica	388	Latin America and the Caribbean	5	2	0	
2011	443	Mexico	484	Latin America and the Caribbean	2	0	1	

[Click on the image or this text to see how this task is carried out in an animated sequence of images.](#)

As you can see in the example, you need to consider what is a duplicate record by deciding which columns you want to compare. If you need additional information, please read the [Excel help page on this topic](#).

Column formatting:

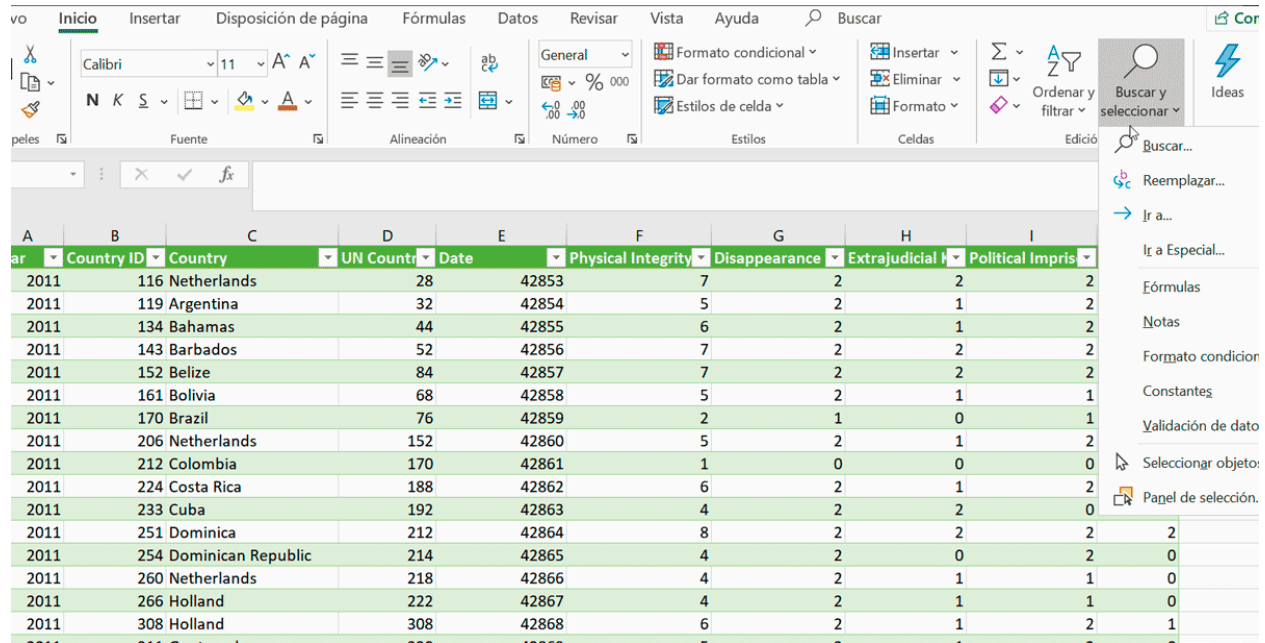


Year	Country ID	Country	UN Country	Date	UN Region	Physical Integrity	Disappearance	Extrajudicial
2011	116	Antigua and Barbuda	28	42853	Latin America and the Caribbean	7	2	
2011	119	Argentina	32	42854	Latin America and the Caribbean	5	2	
2011	134	Bahamas	44	42855	Latin America and the Caribbean	6	2	
2011	143	Barbados	52	42856	Latin America and the Caribbean	7	2	
2011	152	Belize	84	42857	Latin America and the Caribbean	7	2	
2011	161	Bolivia	68	42858	Latin America and the Caribbean	5	2	
2011	170	Brazil	76	42859	Latin America and the Caribbean	2	1	
2011	206	Chile	152	42860	Latin America and the Caribbean	5	2	
2011	212	Colombia	170	42861	Latin America and the Caribbean	1	0	
2011	224	Costa Rica	188	42862	Latin America and the Caribbean	6	2	

[Click on the image or this text to see how this task is carried out in an animated sequence of images.](#)

Each column should have its own type to be displayed correctly. In Excel and LibreOfficeCalc, dates are stored as correlative numbers starting in January 1,1900 (number 1). If you don't select *date* as data type, you will see the correlative number instead of the date.

Name normalization:

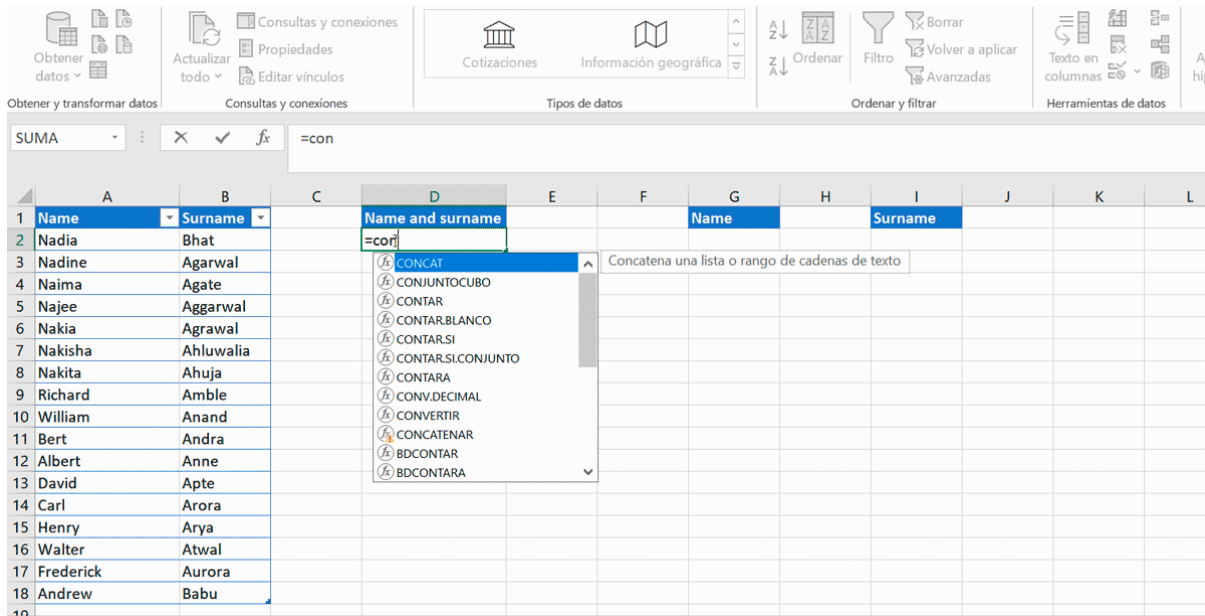


Year	Country ID	Country	UN Country	Date	Physical Integrity	Disappearance	Extrajudicial	Political Imprisonment
2011	116	Netherlands		28	42853	7	2	2
2011	119	Argentina		32	42854	5	2	1
2011	134	Bahamas		44	42855	6	2	1
2011	143	Barbados		52	42856	7	2	2
2011	152	Belize		84	42857	7	2	2
2011	161	Bolivia		68	42858	5	2	1
2011	170	Brazil		76	42859	2	1	0
2011	206	Netherlands		152	42860	5	2	1
2011	212	Colombia		170	42861	1	0	0
2011	224	Costa Rica		188	42862	6	2	1
2011	233	Cuba		192	42863	4	2	2
2011	251	Dominica		212	42864	8	2	2
2011	254	Dominican Republic		214	42865	4	2	0
2011	260	Netherlands		218	42866	4	2	1
2011	266	Holland		222	42867	4	2	1
2011	308	Holland		308	42868	6	2	1

[Click on the image or this text to see how this task is carried out in an animated sequence of images.](#)

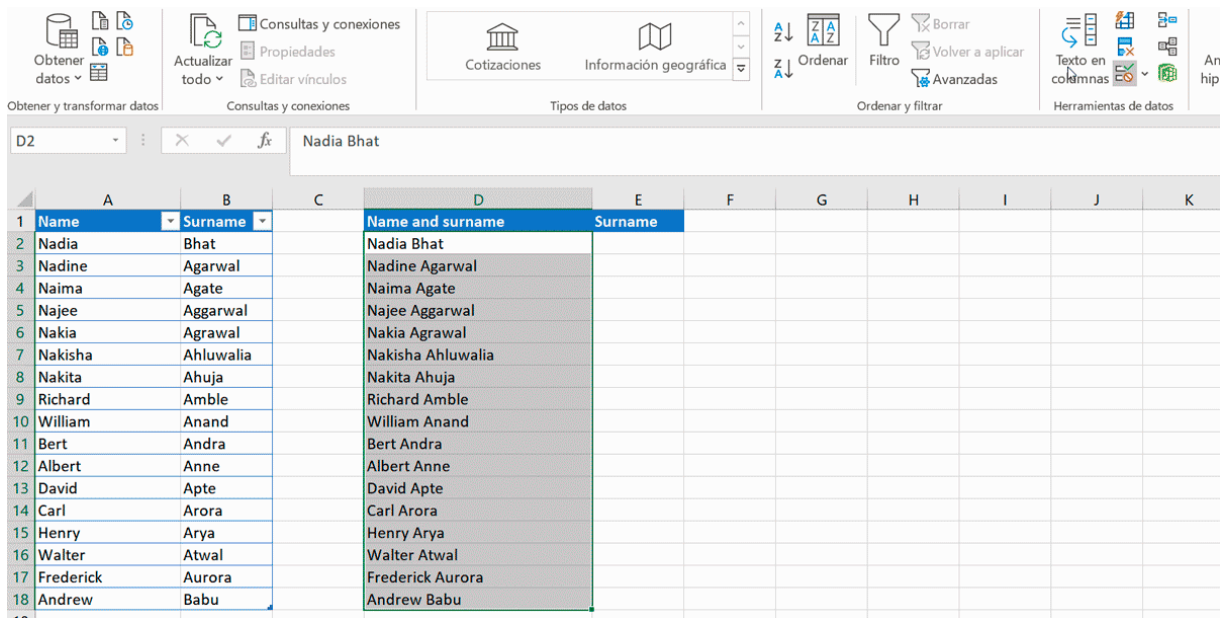
If different people are working in the same database, it is common to have various names to design the same entity. In the example, we use the *find and replace* function to normalize the name of a country.

Merging columns:



[Click on the image or this text to see how this task is carried out in an animated sequence of images.](#)

Separating columns:



[Click on the image or this text to see how this task is carried out in an animated sequence of images.](#)

If you want to discover more about reducing errors in spreadsheets by cleaning data, you can take [this “School of Data” course](#) and you will learn about finding and removing pieces of data, formatting data for different tools, dealing with inconsistencies, and structuring data.

Validation and testing

Now, it's time to ensure that the data is valid according to the rules defined for the new database. Your validation rules and constraints will determine what you should check, but here we present some of the most common actions:

- Data type validation: the data type determines the characteristics of the values that can be stored. For example: integer (positive and negative whole numbers), positive integer, alphanumeric (combination of text and numbers), real (positive or negative decimal numbers), etc.
- Value range validation: you can set a range of valid values. For example, in the *age* field, the database only allows numbers between 0 and 120.
- Mandatory field validation: if your new database includes mandatory fields, check for empty slots in those fields.
- Constraint validation: some fields must include certain characters or a combination of characters. The email *field*, for instance, must contain @.

Ensuring that the data meets these criteria will prevent some records from not being loaded in the new system. You can perform simple validation tests with your spreadsheet software. Read these instructions for [Excel](#) and [LibreOffice Calc](#).

As we mentioned, data migration entails many potential risks, so testing with smaller sets of data is a good strategy before loading. These tests should be conducted with random samples to verify the data quality aspects as well as the correct matching between the source and the target database (the data mapping process, previously explained).

Data loading

The data is ready to be loaded into its new warehouse. Just be sure that, in addition to the syntax rules that we have validated, your data meets other requirements that some databases include.

Uwazi, the free open-source software developed by HURIDOCS for organising, analysing and publishing information, has [specific requirements on how to do this process](#). These requirements include guidelines about how to name and upload files, allowed formats, matching field names, etc.

If your dataset is in compliance with all the requirements, you can load it to the target system.

3. After the migration

Once the data is migrated, it is still required to verify if everything went well. In that case, we will be ready to shut down the legacy system. If something went really wrong, we will still have the option to restore our backup.

The reconciliation test compares the data in the target database against the original source, including aspects such as quantity (checking for missing records and values, or duplicate records), data profiling metrics, structure (relationships established between fields or tables in the old and the new databases), etc. You can do this process manually (with the data profiling and data validation techniques that we have shown previously), or with the functionalities included in software specifically dedicated to data migration.

After the verification and a reasonable period of time, we can wipe out the source database and the migration process will be completed.