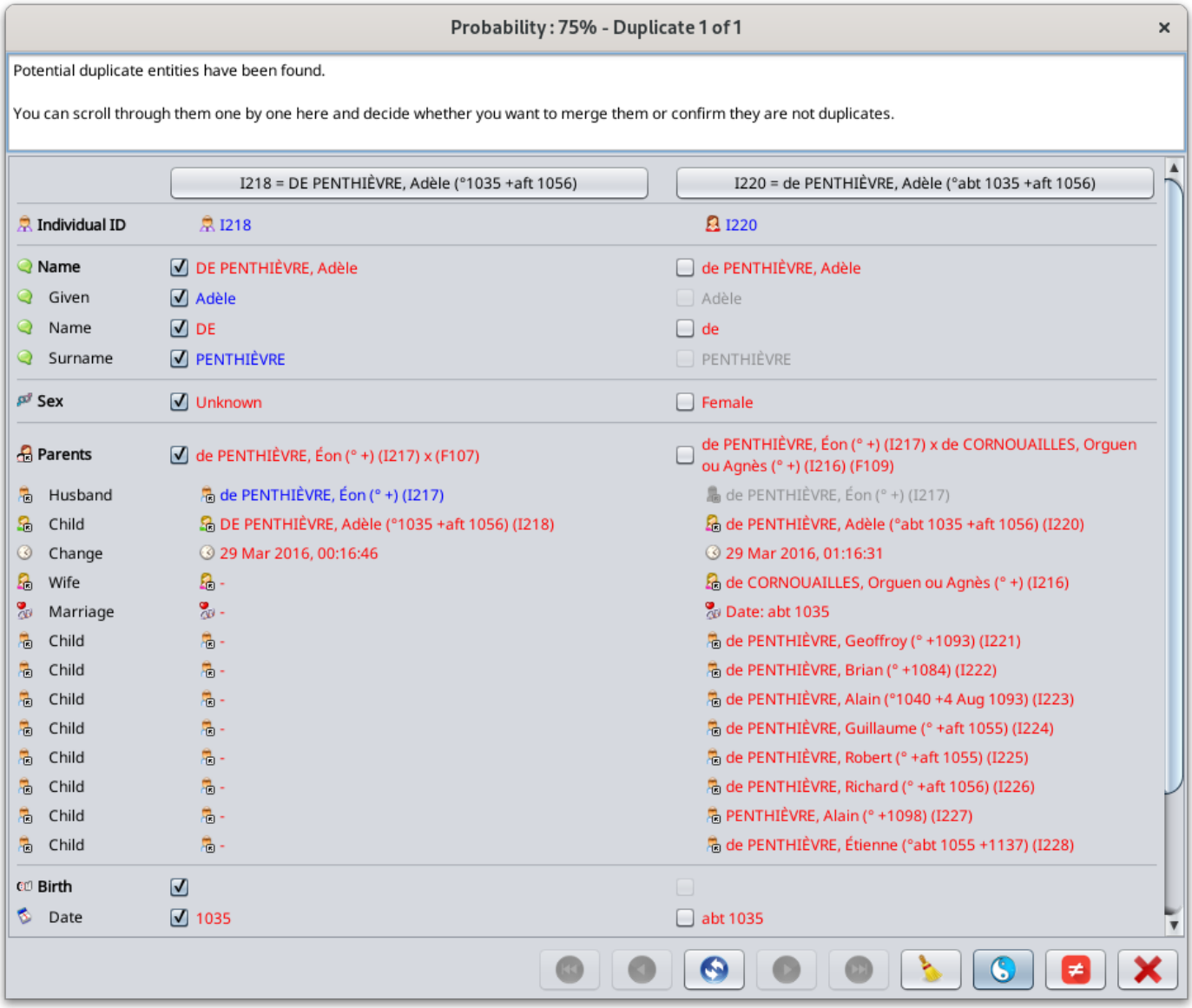
Merge duplicates

This tool detects duplicate entities and allows you to merge them if they are real duplicates, or confirm them as non duplicates if they are not.

There are 3 ways to search and merge duplicates. They are explained in this page.

- **Global search** across the whole genealogy file
- **Automatic detection** each time an entity is modified
- **Manual merge action** from the user to force the merge of two selected entities

The following window is used in all 3 situations. This page describes the components of the tool and how to use it.



Description

The Merge Duplicates tool is made of 3 components.

- **Detection criteria window**: this is where you can set some of the criteria which will determine if two entities should be considered duplicates or not.
- **Merge window**: this is where you will see all potential duplicate entities and decide where or not to merge them.
- **Special note for all non duplicates**: this is where Ancestris stores all your confirmations that two entities are not duplicates.

Detection criteria window

Probability indicator

It is difficult to assess with 100% certainty that two entities are duplicates or are not duplicates. Even a human being can sometimes have difficulty certifying that two individuals or entities are the certainly the same or certainly not the same.

Of course, it would be easier to limit the detection to saying that two individuals with exactly the same surname, first name and date of birth are duplicates. In reality, dates could be either missing or be approximate, first names could be in different order or incomplete, etc. In these cases, you would still want Ancestris to be able to detect something.

Therefore **Ancestris uses a probability indicator**. The more some information is similar, the more probable the entities are duplicates.

Ancestris then lists the potential duplicates according to this indicator in an intent to tell you:" ***While this is not certain, given the similarities in the information between these two individuals, they might be duplicates. And this is the level of confidence that they are***". Then it's up to you to decide of whether merging or discarding the resemblance.

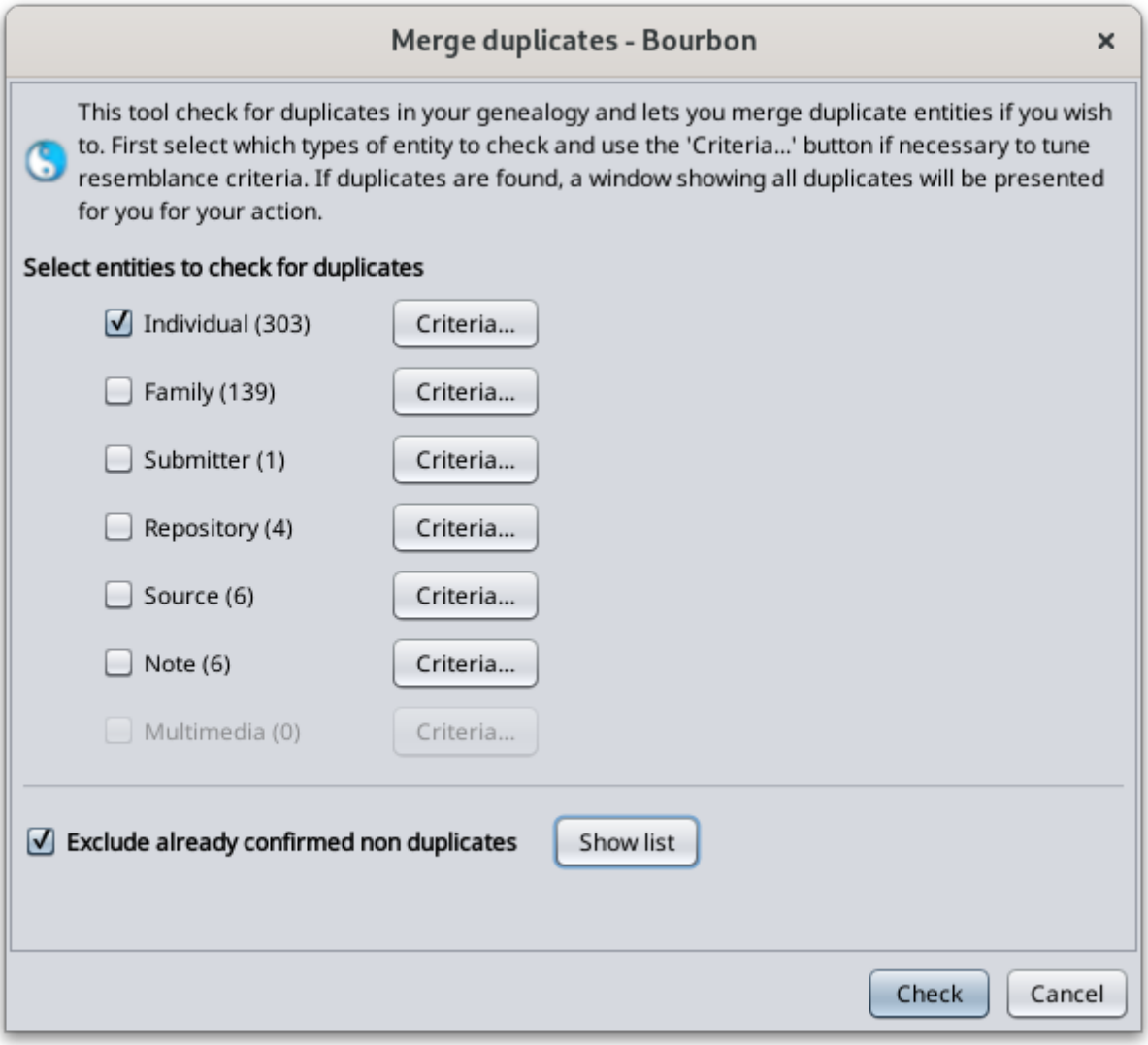
> What this mean is that Ancestris might show you potential duplicates that you will consider as non duplicates, and reversely, Ancestris might miss to show you some duplicates that could actually be real duplicates according to you.
> Please accept our apologies if the detector is not perfect and please let us know if you find such instances.

To calculate this probably indicator, Ancestris uses some criteria.

In the next section you are given the possibility to change some of them.

Criteria selector

When a global search for duplicates is launched, the detection criteria window is displayed with a number of criteria.



Check the entity boxes for which you want to search for duplicates.

Only the boxes of entities that are present in the Gedcom file are enabled. In the example above, as there are no media entities, the corresponding Criteria button is unavailable.

Then you can press the "Criteria..." button to specify some of the criteria for each category of entity.

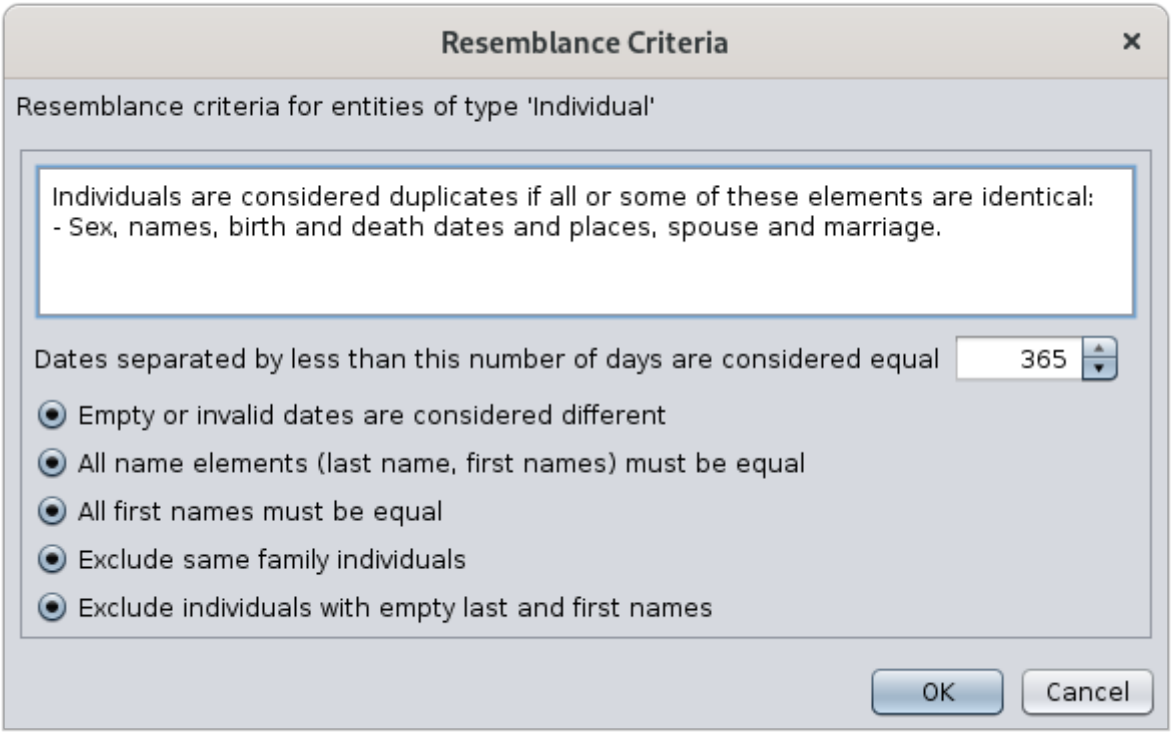
In case you want to exclude known non duplicates from the search, i.e. those that you have already confirmed as non duplicates, check the corresponding box.

To see the list of pairs of entities that you have confirmed as non duplicates, **press the "Show List" button.**

If you press the Show List button, the **list of Non Duplicates** appear.

Entity criteria

The most sophisticated criteria are those of individuals. Here they are.



The criteria are as follows.

The criteria for other entities are either a sub-part of these criteria or are not modifiable.

Identical dates

When are two dates considered identical? When their difference in number of days is close or zero.

If you indicate 365 days for example, i.e. 1 year, two dates will be equal if their difference is less than a year.

If you indicate 30 days, two dates will be equal if they differ by less than a month.

Empty or invalid dates

If a known date is compared to an unknown date, Ancestris will consider them different.

Name elements

Forces all elements of the name to be identical. Conversely, can be identical if only some elements of the name are identical.

First names

Forces all first names to be identical. Conversely, can be identical if only some first names are identical.

Exclusion of individuals from the same family

Individuals from the same sibling or parent-child relationship are not compared.

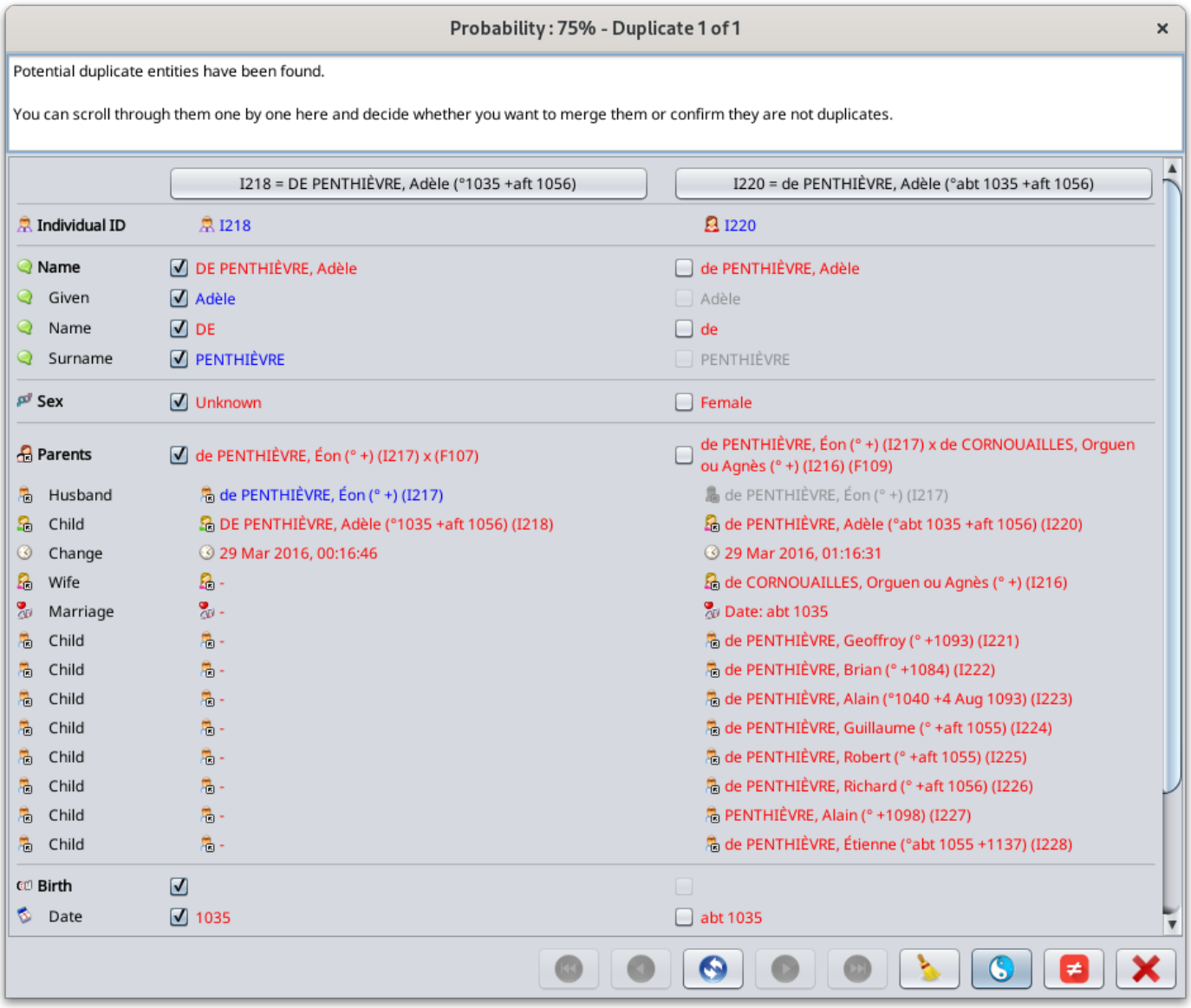
Exclusion of individuals without first or last name

Individuals without first or last names are not compared.

Merge window

For entities compared as duplicates, the following window is used.

Window



This window displays one by one the total **list of all potential duplicates where the probability is greater than 50%**.
The list is sorted **from the most certain pair of duplicates to the least certain** pair of duplicates, by category of entity. For each pair of similar entities, Ancestris gives you a probability percentage.

The **title** of the window indicates the duplicate pair number displayed and the confidence that the two entities of this pair are in fact the same, and therefore to be merged.

A **general message** and displayed and depends on the situation the tool was launched, as global search, automatic detection, or manual action.

Each pair of duplicates made of two entities is displayed in the two columns.

As title of each column, a button allows you to select each of the entities in the editors for more details.

In each column are displayed the properties of each entities of the supposed duplicate.

Values that are different are displayed in red.

Values that are identical are displayed in blue for the left hand side entity, in grey for the right hand side entity.

> The purpose of the comparison is to merge the right entity into the left one if you confirm them as duplicate

Therefore, a check boxe is available for each property on the right hand side to tell Ancestris to keep both selected information of each entity after merging them.

The buttons are used to navigate within the list of duplicate pairs, postpone the decision, merge them now or confirm now as non duplicate.

Button bar

Go to first duplicate Button

Displays the first duplicate of the list in the order of the confidence index, i.e. the most likely duplicate, or the duplicate which is 50 positions before the current one in case there are more than 50 duplicates in the list.

Go to previous duplicate Button

Displays the previous duplicate.

Swap Left and Right Entities Button

Swap the left and right entities in order to merge the two entities on the right one. This is useful if most of the information to be kept after the merge is on the right hand side.

Go to next duplicate Button

Displays the next duplicate.

Go to last duplicate Button

Displays the last duplicate of the list in the confidence index, therefore the least likely duplicate, or the duplicate which is 50 positions after the current one in case there are more than 50 duplicates in the list.

Remove duplicate Button

Removes the potential duplicate from the displayed list.

It is useful if you do not know yet whether the two entities are duplicates or not, and you want to **postpone the decision**.

If a new global duplicate search is started, the duplicate will reappear.

Merge Button

By clicking the Merge button, the **entities will be merged**.

The entity on the right is removed from the Gedcom file and the information which check box is checked on the right hand side is added to the entity on the left.

For information that can only exist once (e.g. birth), it is only possible to keep the information from one of the two entities.

As soon as the merge is done, the window displays the same duplicate with the result of the merge so that you can check that everything has been kept as you wanted.

You can then move on to the next duplicate.

Non duplicate Button

This button **excludes the pair of entities from the potential duplicates**.

It marks the entity pair to be non duplicate and stores this confirmation in the **special "non duplicates" note**.

Close Button

Closes the window.

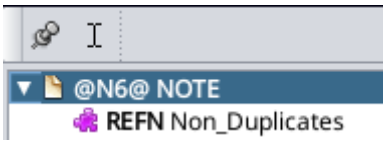
Special note for all non duplicates

A special note is created and updated in Ancestris to store the non duplicate confirmations.

This note stores user confirmations of similar pairs that are actually not duplicates according to you.

It avoids Ancestris detecting them again and again each time the global search or the automatic detection is run.

This note has a reference name called "Non_Duplicates".



The note is updated each time you press the Non Duplicate button in the **Merge window** or update the list using the **Criteria window**.

> We have chosen to store this user information in the Gedcom file itself because we wanted your efforts to analyse the entities and decide that they are not duplicates as a valuable piece of genealogy information that has to be kept and transferred as part of the Gedcom file.
> The Gedcom standard does not cater for this need, hence Ancestris choice to store it this way, in one single note.
> Should the Gedcom standard evolve, such as a NOALIAS tag, we might change the way this information would be managed.

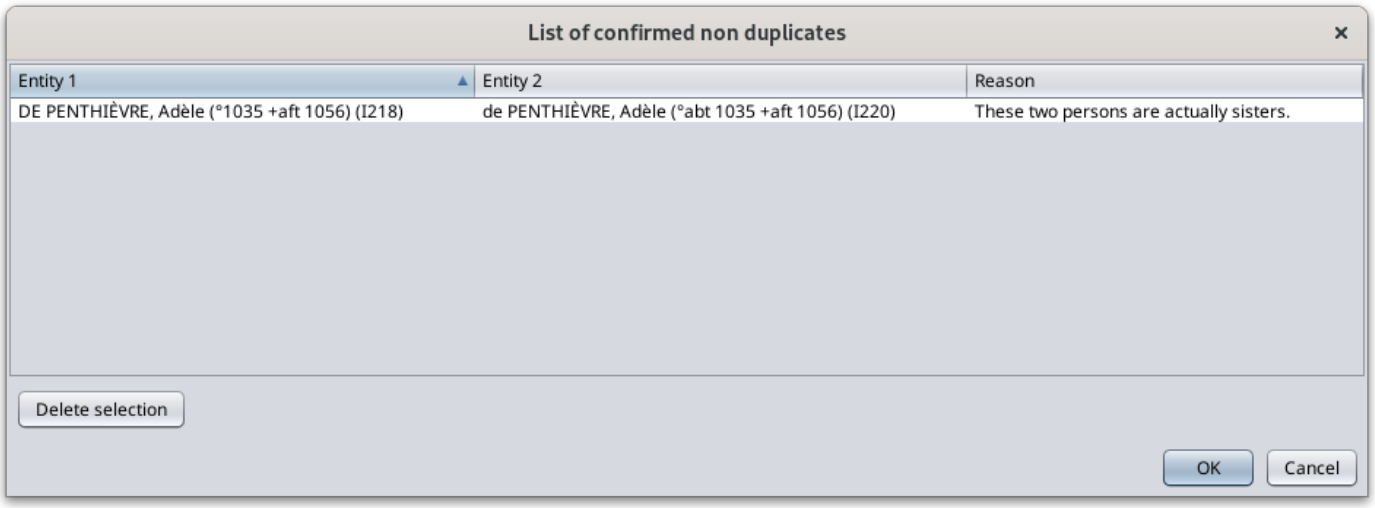
You can see the list of confirmed non duplicates from the **criteria window**.

List of non duplicates

You can see the list of non duplicates by pressing the **"Show list" button on the Detection Criteria window**.

You can sort the lines to find the entities you are interested in.

You can select one or several lines and remove them from the list if you need to.



Usage

As mentioned above, there are 3 ways to use this tool.

- **Global search** across the whole genealogy file
- **Automatic detection** each time an entity is modified
- **Manual merge action** from the user to force the merge of two selected entities

Global search

> The purpose of the global search is both to **identify duplicates throughout the whole genealogy and act on them**, that is decide one by one what you want to do with each them.

Your decision for each duplicate will then be to either

1. **merge** the duplicate,
2. declare it as a **non duplicate**,
3. or **postpone** the decision to later.

You can launch the global search from the Ancestris tools menu.

The duplicate merge tool works in two steps.

- First you specify the **detection criteria** in the corresponding window,

- then you choose how to **merge duplicates** in the Merge window.

While using the tool, the genealogy is changed accordingly

1. Entities you decide to merge are merged with the information you specified to keep
2. Entities you declare as non duplicates are logged into the **special note**.

A message lets you know when you close the merge window.

Automatic detection

> The purpose of the Automatic Detection is to alert you in case the entity you are currently creating or modifying is a potential duplicate of another entity already existing in your genealogy.

The automatic detection of duplicates is activated by default in the **Ancestris preferences**.

As soon as you validate your entry in one of the editors, and if the corresponding preference box is checked, the detection automatically searches potential duplicates of the entity being edited.

All potential duplicates are then presented in the **Merge window**, for you to decide what you want to do with these duplicates.

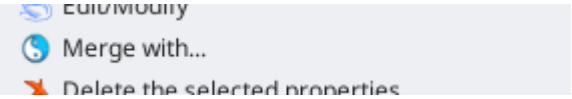
> In the case of the Cygnus editor where several entities can be edited at the same time, all modified entities are checked and therefore Ancestris will list in the Merge window all the potential duplicates of all the modified entities.

Manual Merge action

> The purpose of the Manual Merge action is to merge two entities, regardless of whether Ancestris detected them or not.

> Another purpose is to identify any duplicate for a given entity.

This action is accessible from the **Context menu** on the current entity you want to merge with another one.



When this action menu item is selected, Ancestris asks you which other entity the current entity is to be merged with.

- **To really merge with another entity**, pick the one you think is a duplicate.
- **To just know whether the current entity has got duplicates in the genealogy**, just choose any entity from the list of entities.

Then Ancestris displays the Merge window with a list of potential duplicates among which will be the pair of two entities you chose and its corresponding probability of being the same entity, which can be 0%.

The list will be sorted in decreasing probability and the shown duplicate pair will be your chosen pair of entities.

If the current entity you started from has no other found duplicates, only the pair of chosen entities will be shown in the list.

Then you may decide to merge the current entity you started from with the other chosen entity, or any other entity from the list.

Customization

There are 3 customization settings for the Merge tool.

- The **search entity criteria**, stored in the **User Directory**.
- The **automatic detection flag**, stored in the **Preferences in the User Directory**.
- The **non duplicate pairs**, stored in a special note.

Revision #23
Created 24 May 2020 17:54:43 by frederic
Updated 14 April 2024 20:17:26 by frederic