Step-by-step guide to data clean-up and deduplication

Version 1.5.1 (2021-03-11)

The guide outlines the steps to configure the GPG tool and use the Data Clean-up + ID Generation service.

Contents

Introduction

The following instructions refer to the inter-agency assistance coordination modalities and technical solution agreed by the participating organizations and described in the PDF file included in the shared ZIP folder, which includes the background, the rationale of the chosen approach as well as several technical aspects. The purpose of this document is limited to describing the practical operating procedures related to the data cleaning and deduplication.

Note: The document describes how to operate the data cleanup and deduplication activity in the DAT platform, which aims to calculate also the USCADI codes for the follow-on assistance coordination through Building Blocks. Organizations which decided to perform these technical activities internally, i.e. without uploading beneficiary personal data to WFP's DAT platform don't need to refer to this document.

Data Clean-up & ID Generation Service Changelog

2021-02-11 (current)	Absence of USCADI generated for any input error.
2020-11-26	Row-level validation for USCADI validation.
2020-11-19	Enhanced labels and error messages.
2020-11-17	Tagging logic instead of metadata manipulation for GPG-encrypted files.
2020-10-23	Initial version.

How your data is treated

The data clean up, deduplication and ID generation service is provided by the DAT platform. Registered users are able to utilise the service through the Building Blocks Lebanon user interface.

Uploaded files are stored on the DAT platform for the express purpose of data cleanup, deduplication and ID generation. Only files encrypted with GPG are accepted. The process of data cleanup, deduplication related processing and ID generation are performed automatically with no manual human intervention. The DAT platform automatically produces an *output* data file for the user to download. This file is also GPG encrypted, and only the uploading user, through their private GPG identity, may decrypt it – no WFP staff can decrypt this file.

For troubleshooting purposes, and if necessary, the DAT team will seek the permission of the data owner through the registered user in order to access the contents of an *uploaded* data file, within the data retention period.

The service provided by the DAT platform implements a data retention period of 24 hours, allowing users to download their outputs generated by the platform. Both the uploaded data file and generated output data file are automatically flagged for deletion 24 hours after they are created or stored[1].

All files stored in the DAT platform are encrypted at-rest with server-side encryption.

Process steps

Preparation and encryption of the file

The platform only accepts Excel files that comply with the template file shared ("*Data Submission Template.xlsx*"), please refer to the template file to prepare the document and please ensure it is filled up with data required to generate at least one USCADI code (refer to the table in Annex I): the bare minimum is:

- First Name (Arabic + English)
- Last Name (Arabic + English)
- DOB (YYYY)

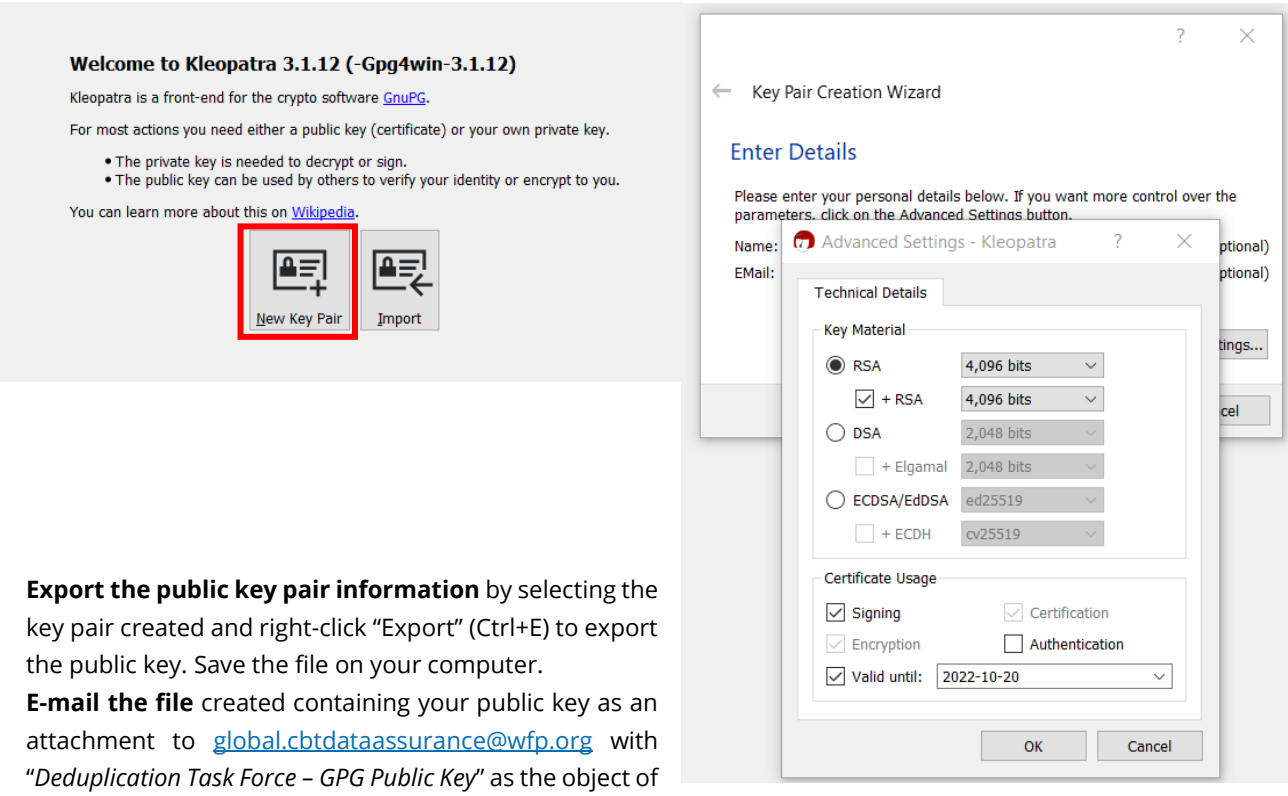
Before the submission of the file into the platform, it is mandatory to encrypt the file using the GPG encryption standard (see steps below).

Note: The steps described below show how to encrypt and decrypt files using the software "Kleopatra", open source and free on Windows and Linux. If you are a Mac OS user, you can download GPG Keychain, free as well.

[1] Due to technical limitations, actual deletion may take up to 24 hours to occur, meaning all data files are automatically deleted up to a maximum of 48 hours after they are created or stored.

To encrypt the file using the GPG standard:

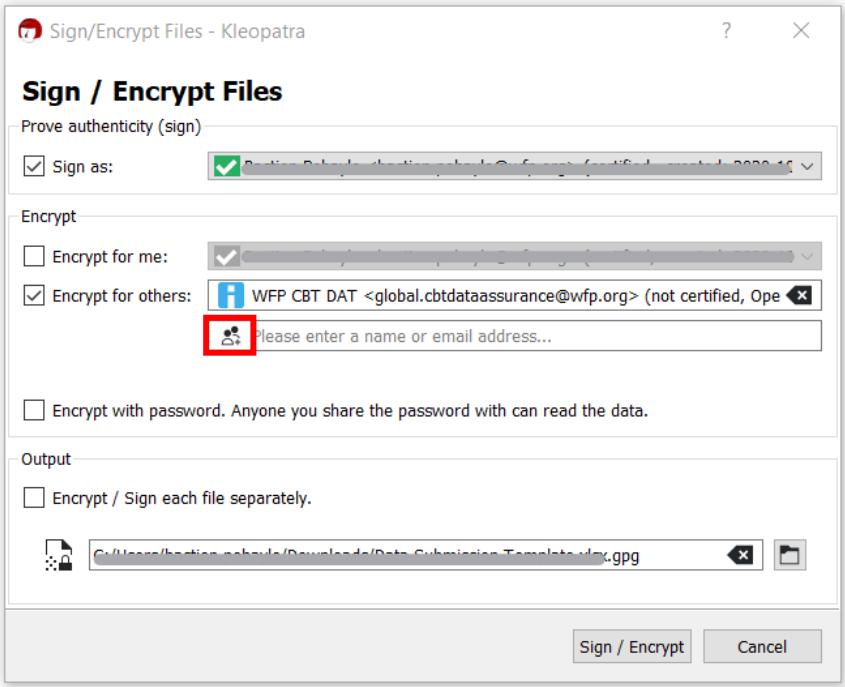
1. **Download** Kleopatra and install it on your computer (click "Next" twice and "Install")
2. **Create a new key pair** by clicking on the icon or "File" -> "New key pair". Enter your e-mail address and name. In the advanced settings, select RSA, check + RSA and select a key length of 4096 bits for both. Make sure the expiration date of your key is at least 1 year.
3. Close "Ok" to close the Advanced Settings window then click "Create".
4. Enter a passphrase as requested, this is your personal password for when encrypting files. Click "Ok" when completed.



5. **Export the public key pair information** by selecting the key pair created and right-click "Export" (Ctrl+E) to export the public key. Save the file on your computer.
6. **E-mail the file** created containing your public key as an attachment to global.cbtdataassurance@wfp.org with "*Deduplication Task Force – GPG Public Key*" as the object of the e-mail. You will receive WFP DAT's public key in return within the next 24 business hours.
7. **Import the public key received by e-mail**: click on "File -> Import" to import the key pair. A window box suggests certifying the public key imported (i.e. an electronic document used to prove the ownership of a public key). If you wish to not proceed to the certification or to do it later, you may disregard the message by clicking "No".[2]

[2] Although not mandatory, a fingerprint verification to certify the public key is recommended. Please e-mail us to global.cbtdataassurance@wfp.org for more information.

8. **Encrypt the file to upload**: Click on the icon "Sign/Encrypt" and select the file to encrypt on your computer by signing with your personal key pair. Encrypt the file for the CBT Data Assurance key specifically by clicking on the icon right to "Encrypt for others" and select "WFP CTB DAT" from the list displayed. Do **not** encrypt with a password.

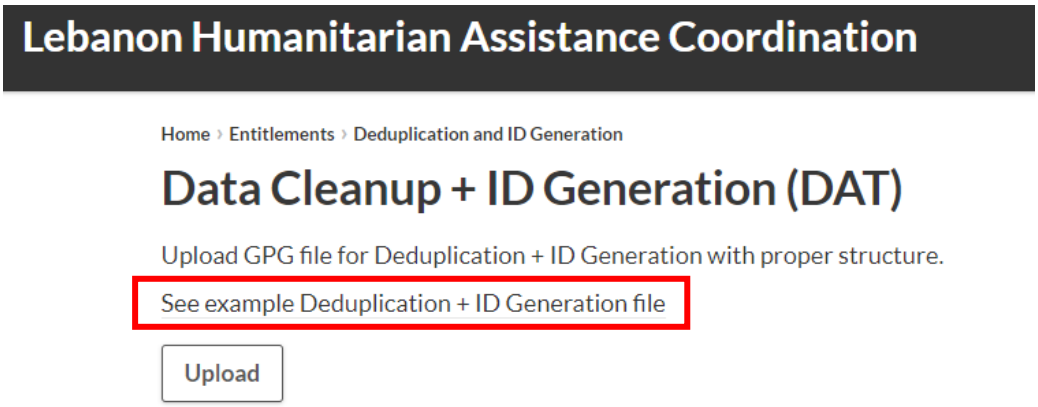


Notes: Your public key must be shared before sending the first file, otherwise the submitted file will return an error, as the platform will not be able to decrypt it.

In case of any updates to your GPG keys, please promptly share the updated public key with the DAT service, otherwise the platform won't be able to decrypt and therefore process your newly submitted files.

Input file

The input file can be downloaded from the platform (https://buildingblocks.lbn.wfp.org/), in the deduplication section. Click on "See example Deduplication + ID Generation file" to download the template file.



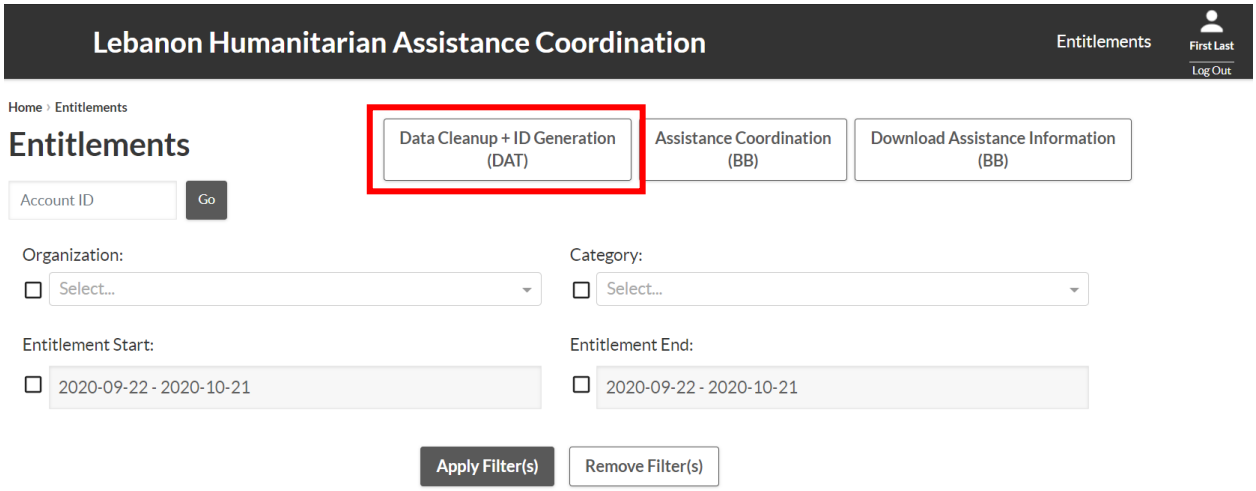
The file contains mandatory fields (red columns) and other optional fields (grey columns). The optional columns field names from the template are for suggestions purposes only, these can be replaced by any other type of data the user wants to add, and with different column names. Column names and data from the input file will remain unchanged in the output file.

If the data is not available for one of the mandatory fields, you can either choose to remove the column or to leave it empty. If you leave it empty, associated error messages will be generated in the output file. In both cases, USCADIs corresponding to the fields available will be generated.

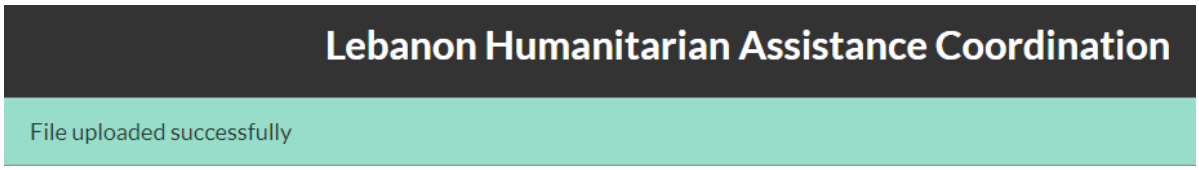
We suggest creating two separate folders "Input files" and "Output files" on your computer to make the distinction between the encrypted input files to be uploaded, and the encrypted output files generated and downloaded from the platform.

Upload of the input file

1. From your browser, access the platform at this link: https://buildingblocks.lbn.wfp.org/[3]
2. Use the credentials that you should have received upon the submission of the access form[4] to login.
3. On the homepage, select the button "Data Cleanup + ID Generation (DAT)"
4. Click on "Upload" to submit the file. Make sure that you upload the encrypted version of the file (the extension of the file is ".gpg)



If the file is correctly uploaded, a green ribbon indicating the successful upload will be displayed:



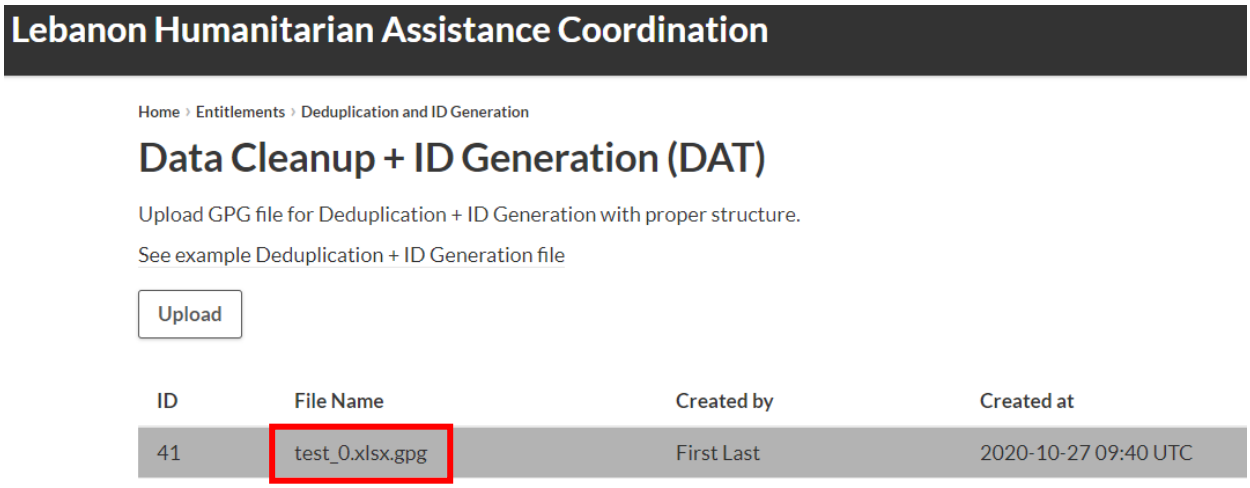
[3] Current supported browsers include Google Chrome and Microsoft Edge

[4] C.f. documentation "Access to the platform and assistance coordination instructions"

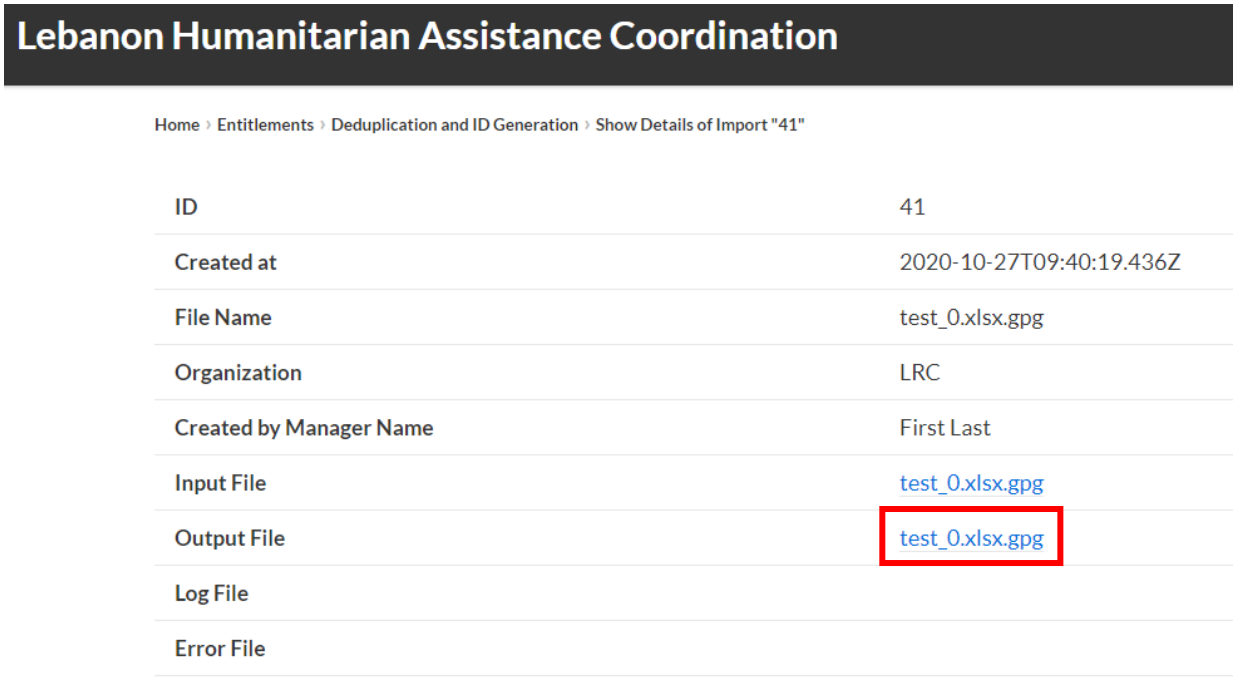
Get back enriched file (output file)

Within a couple of minutes, the platform will process the incoming file. An enriched file will be available for you to download:

1. Click on the name of the file uploaded



2. Click on the file from the "Output File" line to download the output file. The file is now available on your local computer "Downloads" folder.



Note that the retention policy in place keeps both input and output files between 24h and 48h.

Decrypt the output file

After having downloaded the output file:

1. Open the GPG tool (Kleopatra)
2. Click on the icon "Decrypt/Verify" on the top left of the interface
3. Select the output file downloaded from the platform and click open.
4. Select the path where to save the decrypted file. The same folder as the encrypted file is selected by default, we suggest instead to create a new folder for each decrypted output. Click on "Save all".
5. The prompt indicates when the decryption is complete. The decrypted file is available in the same folder as the encrypted file.

Understand the enriched file

In case of successful processing (otherwise see the "Errors and Troubleshooting" paragraph), the file you receive back contains the same number of rows you submitted; the following changes are applied to the content:

- Clean-up: The cleaning and standardization of names is important to make sure that no misleading or redundant information is introduced. The data cleanup process consists of the removal of punctuation, spaces, non-letter characters (including digits) from names. Birthdates are formatted in a consistent way. Arabic character names are cleaned such that characters associated with typographic variations are excluded. These include dropping of *hamzas*, dropping the *madda* from the *aleph* and removal of diacritics. Arabic language expertise is required with the latter.
- Extra columns for deduplication: The output file adds 20 extra columns to the original file, 12 columns representing the detailed deduplication work from the algorithm and 8 columns representing the deduplication identifiers generated.
- USCADIs are only generated when the columns required are present and correctly parsed. All USCADI columns will be empty if any error in one of the input columns (c.f. "Errors" column for more detail).

<table>
<tr><th>Output field</th><th>Description</th></tr>
<tr><td>Duplicate confidence type</td><td>Ranked scale from "Low" to "Identical" and indicating the confidence level about the duplicates flagged by the algorithm.
<table>
<tr><th>Confidence scale</th><th>Description</th></tr>
<tr><td>Blank</td><td>No duplicate</td></tr>
<tr><td>Raw identical</td><td>The records are identified as identical prior to the cleaning process.</td></tr>
</table></td></tr>
</table>

	Cleaned identical	The records are identified as identical after the cleaning process (but not before it).
	High	The three methods used for the deduplication (c.f. annex 2) agree on the similarity of the records flagged.
	Medium	Two of the three methods used for the deduplication (c.f. annex 2) agree on the similarity of the records flagged.
	Low	One of the three methods used for the deduplication (c.f. annex 2) identifies a similarity of the records flagged.
Raw signature	Signature based on all fields of the file. If available, the record is identical to another record in the file.	
Cleaned signature	Signature based on all fields of the file, after the clean-up. If available, the cleaned record is identical to another record in the file.	
Metaphone signature	Signature based on the Latin characters only. If available, a duplicate is flagged based on the Metaphone methodology.	
Partial letter signature	Signature based on the first 3 letters of each Arabic field. If available, a duplicate is flagged based on the partial letter methodology.	
Soundex signature	Consists of the Soundex algorithm applied to Arabic names only. If available, a duplicate is flagged based on the Soundex algorithm.	
Errors	If an error message is present, the record's fields should be checked. Any deduplication IDs produced for that record should be considered unusable.	
USCADI (*1 to8*)	USCADIs (Unique, Singular, Common, Anonymous Deduplication Identifier) are generated according to the data points available, and therefore on what data fields have been collected (c.f. annex 1).	

Follow-up activities

Once you have successfully received and decrypted the file, you should:

1. **Check for error messages**: If an error message is present, you should review the record according to the reason of the error:
   - Name field is empty
   - DOB (full) is not in the expected date format (YYYYMMDD)
   - Non-Arabic characters in Arabic name field
2. **Identify raw identical duplicates**: Sort the "Raw signature" column to show what row is duplicate of the other. You can then inspect the rows and remove the duplicate.
3. **Identify cleaned identical duplicates**: Sort the "Cleaned signature" column to show what row is duplicate of the other. You can then inspect the rows and remove the duplicate.
4. **Identify duplicates with a high confidence level**: Filter the "Duplicate confidence type" on "High confidence", then identify the rows by filtering on any of the method (Metaphone, Partial name or Soundex).
5. **Identify duplicates with a medium confidence level**: Filter the "Duplicate confidence type" on "Medium confidence", then identify the rows by filtering on the two methods flagging the duplicate.
6. **Identify duplicates with a low confidence level**: Filter the "Duplicate confidence type" on "Low confidence", then identify the rows by filtering on the method flagging the duplicate.

Notes: The deduplication service is a decision-support tool, however the final decision following the reviewing process is in the organization's hands. It has the decision-maker role on keeping or not potential duplicates. Please note that removing duplicates with medium and particularly those with low confidence should be done carefully, as it may lead to exclusion errors.

Errors and troubleshooting

Failed upload of the input file in the platform

A red ribbon with an error message is displayed in case of any unsuccessful upload attempt:

In this case, proceed to the following checks:

- The file uploaded is encrypted (".gpg")
- The file has been encrypted to the right public key (CBT Data Assurance) in the GPG tool
- The format of the input file is correct[5].

Error file and no output file after upload

In case of unexpected errors during the processing of the file uploaded, an "Error file" is available in place of the "Output file" to download. The error file may contain an error log similar to the message:

```
[ErrorHandler]: An unexpected error occurred; the error trace will be written to [FILE NAME].lambda_error.txt
```

In that case, verify the input file for any possible formatting issue or value error. If the file is valid, create an unencrypted sample file reproducing the error, without the sensitive data. Send the file with a description of the issue to global.cbtdataassurance@wfp.org for the WFP Data Assurance Team to follow up.

Errors and empty values in the output file

If any error message in the output file, identify the issue and re-upload a list that does not generate **any error** message before uploading a file for BB Assistance Coordination.

Some error messages may be trivial to address (e.g. date format string), but others may require follow up directly with the beneficiary, e.g. to confirm names or dates of birth.

Please refer to the messages available in the "Errors" column for more context about a missing value. You may also refer to the *Rationale for Error Messages* table (Annex. 3) for the list of possible errors.

[5] The template is available to download on the platform https://buildingblocks.lbn.wfp.org/ under the "Data Cleanup + ID Generation (DAT) section.

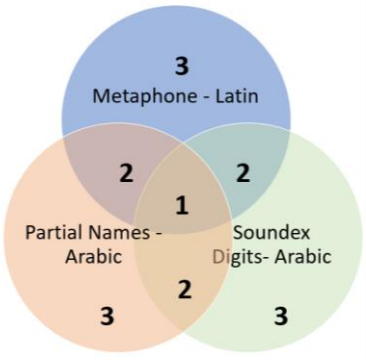
Annexes

Annex 1: USCADI Table

The table below explains which fields are utilised to generate the 8 different USCADIs:

USCADI	Soundex Arabic Phonetic Code				Metaphone I English Phonetic Code				DOB		Salt
#	First Name	Last Name	Father's First Name	Mother's First Name	First Name	Last Name	Father's First Name	Mother's First Name	YYYYMMDD	YYYY	***
1	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
2	✓	✓	✓		✓	✓	✓		✓		✓
3	✓	✓		✓	✓	✓		✓	✓		✓
4	✓	✓			✓	✓			✓		✓
5	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
6	✓	✓	✓		✓	✓	✓			✓	✓
7	✓	✓		✓	✓	✓		✓		✓	✓
8	✓	✓			✓	✓				✓	✓

Annex 2: Venn Diagram of the Three Deduplication Methods



1: High-confidence duplicates

2: Medium-confidence duplicates

3: Low-confidence duplicate

Annex 3: Rationale for Error Messages

INPUT COLUMNS										ERROR MESSAGE
First Name (Latin)	**Father's First Name (Latin)**	**Last Name (Latin)**	**Mother's First Name (Latin)**	**First Name (Arabic)**	**Father's First Name (Arabic)**	**Last Name (Arabic)**	**Mother's First Name (Arabic)**	**DOB (yyyyMMdd)**	**DOB (yyyy)**	**Error Description**
	David	Williams	Amy	ياسر	سامح	التوتنجي	ميرا	19650708	1965	"First Name (Latin)" is mandatory
Brent		Harmon	Melinda	عبد الرّشيد	مياس	الفتياني	بنان	19160410	1916	"Father's First Name (Latin)" is blank
Aaron	Ernest		Robin	نور الدّين	سالم	غطفان	زكية	19180718	1918	"Last Name (Latin)" is mandatory
Jeffrey	Nathan	Stewart		سرحان	سالم	البيسار القعقور	ريتال	19820922	1982	"Mother's First Name (Latin)" is blank
Edward	Frank	Fleming	Ashley		صدّاح	الأزد	بلقيس	19050302	1905	"First Name (Arabic)" is mandatory
James	Norman	Young	Tammy	عزاز		بنو هلال	وفاء	19120216	1912	"Father's First Name (Arabic)" is blank
Jason	Matthew	Roth	Shannon	مُرسي	فرزدق		بنان	20040421	2004	"Last Name (Arabic)" is mandatory
Jake	James	Knight	Bianca	حسني	نافع	الزماميري		19301205	1930	"Mother's First Name (Arabic)" is blank
Jeffrey	Raymond	Davis	Emily	موسى	صهيب	طيء	آيات		1950	"DOB (yyyyMMdd)" is blank
Donald	Kevin	Hill	Kathryn	غامد	زكريا	بنو شعبة	نشوة	19550914		"DOB (yyyy)" is mandatory
a12	Nathan	Olsen	Carolyn	يانع	مرعي	الأيوبي	افتكار	19060812	1906	"First Name (Latin)" does not meet minimum length of 2 after cleaning
Brandon	d!	Day	Lauren	سنان	ناضر	القضماني	ضحى	20020225	2002	"Father's First Name (Latin)" does not meet minimum length of 2 after cleaning
Michael	Ryan	x23	Lauren	حامد	عبد الجليل	بنو الدئل	بشرى	20181127	2018	"Last Name (Latin)" does not meet minimum length of 2 after cleaning
Aaron	David	Thompson	g$$	جاسم	تامر	مراد	غيداء	20020824	2002	"Mother's First Name (Latin)" does not meet minimum length of 2 after cleaning
Aaron	Joseph	Young	Jacqueline	ي12'	سعيد	البيسار القعقور	تالا	19840928	1984	"First Name (Arabic)" does not meet minimum length of 2 after cleaning
John	Kyle	Ashley	Emily	جرير	qي	غنيم	مريم	20071223	2007	"Father's First Name (Arabic)" does not meet minimum length of 2 after cleaning
Kyle	James	Reilly	Megan	عبد السّلام	شاطر	ي!	جود	19720422	1972	"Last Name (Arabic)" does not meet minimum length of 2 after cleaning
James	Arthur	Knapp	Patricia	جدير	خلف	صيداوي	triي	19370518	1937	"Mother's First Name (Arabic)" does not meet minimum length of 2 after cleaning
William	Charles	Tyler	Morgan	رجب	ظهير	شمران	راما	12/16/1970	1913	"DOB (yyyyMMdd)" is not in the required format

Robert	Chris	Evans	Michelle	فضل	هلال	ابو السعود	ربى	19661229	19701216	"DOB (yyyy)" is not in the required format
Jeffrey	Richard	Perez	Jo	سنام	معتوق	اسطمبولي	إلينا	19981022	1!9-."!32	"DOB (yyyy)" is not in the required format
Jeffrey	Zachary	Allen	Emily	شدّاد	مأمون	بتروني	صيداوي	19701123	1971	DOB year mismatch, no USCADIs will be generated.
Kenneth	Justin	Suarez	Lisa	صالح	نظام	جذام	لورين			"DOB (yyyyMMdd)" is blank "DOB (yyyy)" is mandatory No USCADIs can be generated due to missing DOB
Daniel	Thomas	Turner	Sarah	نوح	علاءالدّين	الوعري	وسجايا	20500101	1922	"DOB (yyyyMMdd)" in the future
Christopher	Scott	Holder	Ashley	عزّت	ممتاز	البامية	ريف	19710918	2050	"DOB (yyyy)" in the future
Jason	Jeffrey	Mcclain	Stephanie	محفوظ	خطيب	البغدادي	جودي	20500101	2050	"DOB (yyyyMMdd)" in the future "DOB (yyyy)" in the future No USCADIs can be generated due to missing DOB
Jesse	Mathew	Moore	Doris	عبد العزيز	ربيع	مضر	جودي	18951121		"DOB (yyyy)" is mandatory Warning: suspiciously old "DOB (yyyyMMdd)"
Michael	Andrew	Allison	Amber	رشدي	عبد العليم	شهران	إخلاص		1895	"DOB (yyyyMMdd)" is blank Warning: suspiciously old "DOB (yyyy)"
Joshua	Derek	Doyle	Kristin	راجح	كنان	السادة الراويون	حلا	18951121	1895	Warning: suspiciously old "DOB (yyyyMMdd)" Warning: suspiciously old "DOB (yyyy)"
Keith	Jerome	Davis	Victoria	نعمان	مازن	أكلب	إباء	12/16/1970	2050	"DOB (yyyyMMdd)" is not in the required format "DOB (yyyy)" in the future No USCADIs can be generated due to missing DOB