The Administrative Business Register (ABR) System Description Document

Develop the ABR System Based on the Initial Database (ABR-V1)

TABLE OF CONTENTS

1. Background 2
2. ABR System Objective 3
3. ABR Version One (V1) – Synthetic Presentation 4
4. Roadmap from V1 to a mature System 8
5. The Risks and Difficulties (Challenges) 14
6. Human Resources Capacities needed to manage the ABR System 15
7. ABR linkages management – an overview 16
8. Partners' roles & benefits in the ABR system 19

Appendix 1: Life Cycle of the ABR and Data Flow 21
Appendix 2: Cleaning Specifications 25
Appendix 3: Variables 28
Appendix 4: The developments made for the ABR-V1 that shall be integrated in the ABR System next phase 32
Appendix 5: "Description of Partner's files" 39
Appendix 6: Matching and Validation Process 41
Appendix 7: Database Description 47
Appendix 8: Data Flows 50
Appendix 9: Future System 53
Appendix 10: PowerPoint synthetic presentation of the achievements and perspectives 55
Appendix 11: PowerPoint of the presentation of the ABR at the final event of the project 61

Document compiled by the Palestinian and French Project Team during November 2018 – January 2019

1. Background

The State of Palestine, in building its institutions, intends to have an efficient system of Business registers that should serve its institutions' duties and tasks.

By a Cabinet's decision of 2008, the Central Bureau of Statistics (PCBS) was given the responsibility to set-up and chair a National Committee responsible to develop and maintain a national common Administrative Business Register (ABR) that would be instrumental for allowing each Palestinian institution to manage its own register with the help of the shared information managed inside the ABR system (identification variables).

The mandate of the Committee covers administrative, legal and technical preparations pertaining to the development and maintenance of a national Administrative Business Register (ABR).

To achieve this goal, PCBS conducts a participatory approach that involves all Palestinian institutions represented in the Committee, in particular the Ministry of National Economy, the Ministry of Finance, the Ministry of Local Affairs and the Ministry of Agriculture.

With a French funding and experts' assistance, PCBS has developed a first version of an ABR that is presented here under. The document also describes the next steps that will be carried on in the very next years with the perspective to achieve a mature "ABR System".

2. ABR System Objective

2.1 ABR Objective

1. The presently developed ABR aims at providing an updated register of active economic units including their identification data (names, addresses,…) to all Palestinian Institutions that are in charge of a public service mission (the partners).
2. This ABR System will provide the linkages that have been made between the ABR records and the records Palestinian Institutions register in order to help them with the coverage of their files and the quality of the information they have.

2.2 ABR System basic characteristics

3. The ABR register aims at being exhaustive with no duplicate of establishment or enterprise records. At the present stage, it should be fully updated once a year, some partners providing updated information more frequently that could be used for partial updates.
4. The register contains variables (the "ABR shared variables") that allow contacting economic units. These shared variables are listed in §3.2 below.
5. The ABR tables information has to be provided to each partner in order for him to use it for updating his own register.
6. The register has to be regularly updated (new created establishments added, new linkages with partners' records provided, new values for shared variables made available). These updates rely on the information the partners provide to this aim.
7. The ABR tables connect the establishments registered to the enterprise they belong to.
8. The partners' records are linked with ABR enterprise and establishments records in order for each partner to know what he has\has not yet registered (or not yet provided enough information for a link to be made).
9. The ABR System has to store all the ABR Data in efficient, reliable, and secured way to make it useful to its users.
10. The ABR System shall provide at a later stage gateways for each partner to access ABR shared variables and their own provided data on line. At this moment, partners will upload their updated records in order to fill in the ABR System and will be able to download ABR shared variables' data. A friendly Graphical user interface (GUI) will enable users to access the System easily.

3. ABR Version One (V1) – Synthetic Presentation

With a technical and financial support from the French Government, PCBS has developed a first version of a shared Administrative Business Register (ABR-V1) achieved by the end of the year 2018.

The V1 is the first ABR produced with the data provided by MoNE (WB and Gaza), MoFP (VAT), National Chamber of Commerce and ten large municipalities.

It consists of two ABR tables (Enterprise table and Establishment table).

The Enterprise table records the ABR Enterprise IDs that regroup under one ID the Establishments that belong to it (the local units where an activity is carried out).

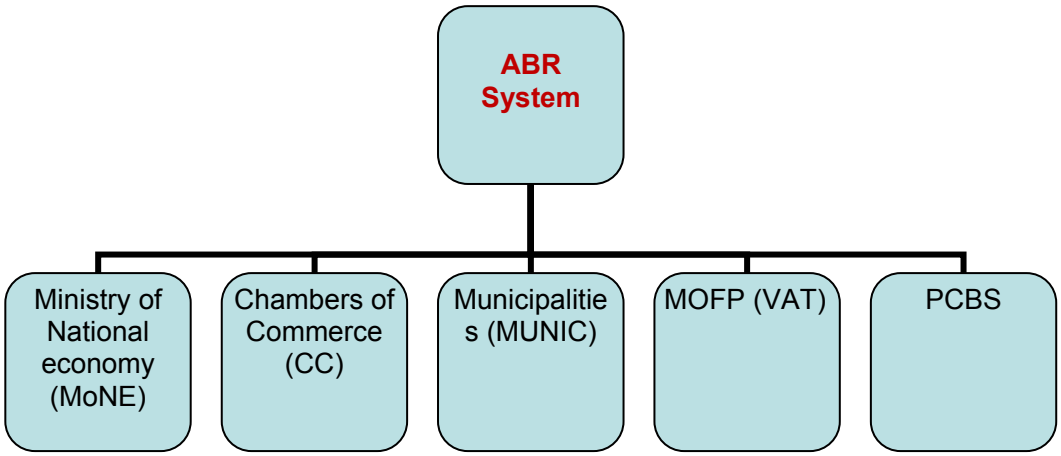
The Establishment table records *inter alia* the best available information from partners files that allow to contact them (owner/commercial name, address…). Each establishment is represented by its ABR Enterprise ID and its specific ABR Establishment ID.

Almost 40 000 units are already registered for the West Bank. This version does not include Gaza (with limited exceptions).

3.1 ABR V1 General Description

The ABR-V1 covers enterprises and establishments (local units) in the West Bank that are registered in the files of at least two institutions and that are in activity. Almost 40 000 establishments (to compare with the 100 000 establishments recorded in the last Census) are already included in this ABR-V1.

This ABR-V1 results from the matching of the information provided by the following institutions (the Partners): MoNE, MOFP (VAT), Chamber of commerce, ten municipalities (Jenin, Tubas, Tulkarm, Nablus, Qalqilia, Salfit, Ramallah, Albireh, Bethlehem, Hebron) and with inputs provided by PCBS.



Each ABR record is connected to the records of the partners that correspond to the same enterprises and establishments.

The present ABR version fulfils crucial functions:

- It registers enterprises and their establishments (local units) deemed active at the moment of being entered in the System.
- It contains shared variables giving basic identification data for each record. These variables are: personal ID or company ID, commercial or owner name, address components (including the GIS coordinates and headquarter for multi establishments enterprises). All variables are not available for each recorded establishment.
- It provides the links with the specific units registered by partners. It should be noted that partners do not necessarily register enterprises or establishments. They may, for example, register companies, taxpayers, establishments' activities…

Each partner has full access to the shared variables of the ABR establishments and enterprises tables and has the links with its own file' records. The partners can check if ABR units that they should cover are missing in their own file. They can also find updated information about owner's data and place of activity addresses.

On this solid base, the next phase of development of the ABR should address the following issues:

- Extend the ABR coverage with more linkages with partner files (V2 – 2019).
- Update the ABR with enterprises and establishments created (V2 – 2019).
- Updating the values taken by the shared variables with the information provided by each partner. (V2 – 2019).
- Extend the data collection to be done by some partners for some variables. The aim is to improve the quality, the freshness, the overall information made available (V3 – 2020). The coverage of the ABR will also make further progress.

Progressing on these improvements will need an active and permanent involvement of the partners for collecting and providing better quality information for some of the variables and for inserting some needed data for a smooth management of the ABR. This expected involvement shall be described in appendixes.

3.2 ABR-V1 Shared Variables and Tables

The ABR **variables** that are **shared** with all partners are deemed non-confidential information from one institution.

The list of variables that can be shared with partners to improve the quality of their data and files are as the following:

- Commercial Name for companies.
- MONE ID: for companies.
- Owner Name for non-companies.
- Personal ID for non-companies.
- Governorate.
- Locality.
- Street.
- Headquarter (*).
- GIS Coordinates.
- Year of creation.

(*) For multi-location enterprises, a headquarter information is proposed corresponding to one of the public addresses. This "headquarter" information in the ABR table has no legal status.

The ABR System includes the following **tables**:

- ABR Establishment table and ABR Enterprise table.
- One table for each partner (MoNE, MOFP-VAT, CC, 10 Municipalities) that contains the variables it has provided and, for each of its records linked to the ABR Establ table, the two ABR IDs that correspond to this ABR record.

The variables provided by a partner are used for (1) linking its records to the ABR records and (2) choosing the best values for the shared variables of the ABR. The algorithms used for taking the best values available in partners' files result from analysing the quality of the values registered by each partner and their availability in each partner's file.

Some examples are provided for describing how the best available values are selected.

ABR Shared Variables Sources Priorities		
variable name	For companies	For non_companies
ABR_Enterprise_ID	generated by the validation system	
ABR_Establishment_ID		
Commercial_ID	MoNE	
Owner_ID (Personal ID)		VAT then Munic then Chamber
Commercial_Name	MoNE	
Owner_Name		VAT then Munic then Chamber
Address (governorate, locality, GIS coordinates)	Census then Munic then VAT then Chamber	
Headquarter or not	Census	
Year of creation	Census (possibly others)	

4. Roadmap from V1 to a mature System

On the road to a mature system, steps have to be followed. The annual updated versions of the ABR should include achieving improvements decided for the given year.

The implementation of aimed improvement will have to rely on “contracts”, “commitments”, “decisions taken and followed” with partners. Depending on each improvement to conduct, the commitments asked from each partner can be specific. Only the sum of the specific commitments allows achieving the aims of the year. A crucial point for success will be the effective involvement of the partners.

4.1 V2 – 2019

V2 should lead to a better coverage of the ABR resulting from partners improving the collection of the variables that are used to match records.

Improvement in registering these variables will allow to increase dramatically the size of the ABR (reaching 70 000 establishments for 2019 is a reasonable objective).

This will strongly be depending on the work done by the partners for the four main matching variables that are:

- MoNE-ID;
- Commercial name;
- Personal-ID;
- Owner’s name.

With these data massively present in the partners’ files, the objective is to introduce units (enterprises and establishments) in the ABR which still have not been included due to the lack of registration of these identifying variables in partners records.

If a high level of registration of these variables is reached, one could reconsider the principle followed for V1, namely of needing the presence of the unit in at least two files in order to introduce an establishment record in the ABR. Having these values well registered should help not being at risk of duplicating an establishment in two separate ABR records even when the unit is not matching with a unit of another partner.

4.2 V3 – 2020

The V3 step will aim at improving the value added of the ABR on two levels:

- Scope covered by the ABR;
- Quality of the values given to the variables (other variables than main matching variables).

Scope coverage by the ABR has been already mentioned during the preparation of the ABR. Some productive activities are not recorded by many partners because of their lack of a physical location (office, shop, …). They have not been included in the ABR-V1.

Taxis represent an example. Obviously, this activity is contributing to the creation of economic value, taxis are deemed paying taxes, and are registered by some institutions. They should be registered in the ABR.

Ahead of making decisions for these kinds of economic activities, it is suggested to look at files as the VAT file to examine units declaring VAT and not present in the ABR V1, and also to analyse the registration by potential partners to the ABR not yet included as partners like the Ministry for Transportation in the case of taxis. If the analyse conducted in 2019 concludes that such an additional field can be included, the V3 should do it.

Other activities not included at present could also be studied.

Quality of the variables' values registered

Some partners are more able to be in regular contact with their registered population and therefore able to ask them updating their information.

Analysing the scope of the information that each partner can be keen to collect with care for quality, "contracts" can be settled with each one for being the provider of updated information for such or such variable.

Also, a partner could be asked to become responsible for collecting high quality data registration for variables it does not collect to date. These demands should correspond to information of common interest to partners' institutions.

Some variables have already been selected that should be collected by a partner able to bring a large coverage of the units that are to be registered in the ABR: headquarter location for multi-establishments enterprises, GIS coordinates, main economic activity, rough number of employees.
It could be expected that municipalities take the responsibility to collect those information on their respective territory.

To achieve an efficient cooperation between PCBS in charge of managing the ABR and municipalities in charge of such data collection, clear conditions should be set that contribute to the expected delivery. It can include the use of tablets that helps homogeneous field work.

At the V3 stage, when partners files' records will include MoNE and Personal IDs, Owners and Companies names, and if GIS coordinates are systematically available thanks to municipalities data collection, the ABR system should be able to solve smoothly the duplications issues.

Preparing such an improvement is an issue for 2019 that could then be implemented in V3-2020.

4.3 2020+ – A mature system

For the mature system developed to be operational from 2020 on-wards, a sustainable ABR System will include all features to cover updates, extending scope ... and needs resulting from an online system of exchanges, with rules defined for each partner.

Different components are to be set:

- Periodicity and dates of the updated data transmission. They should mainly depend on the characteristics of each partner's data collection (at what time of the year do they update their own files, what do they update and at what rhythm). Each partner situation has to be taken in consideration. All partners are informed about the periodicity followed by each one;

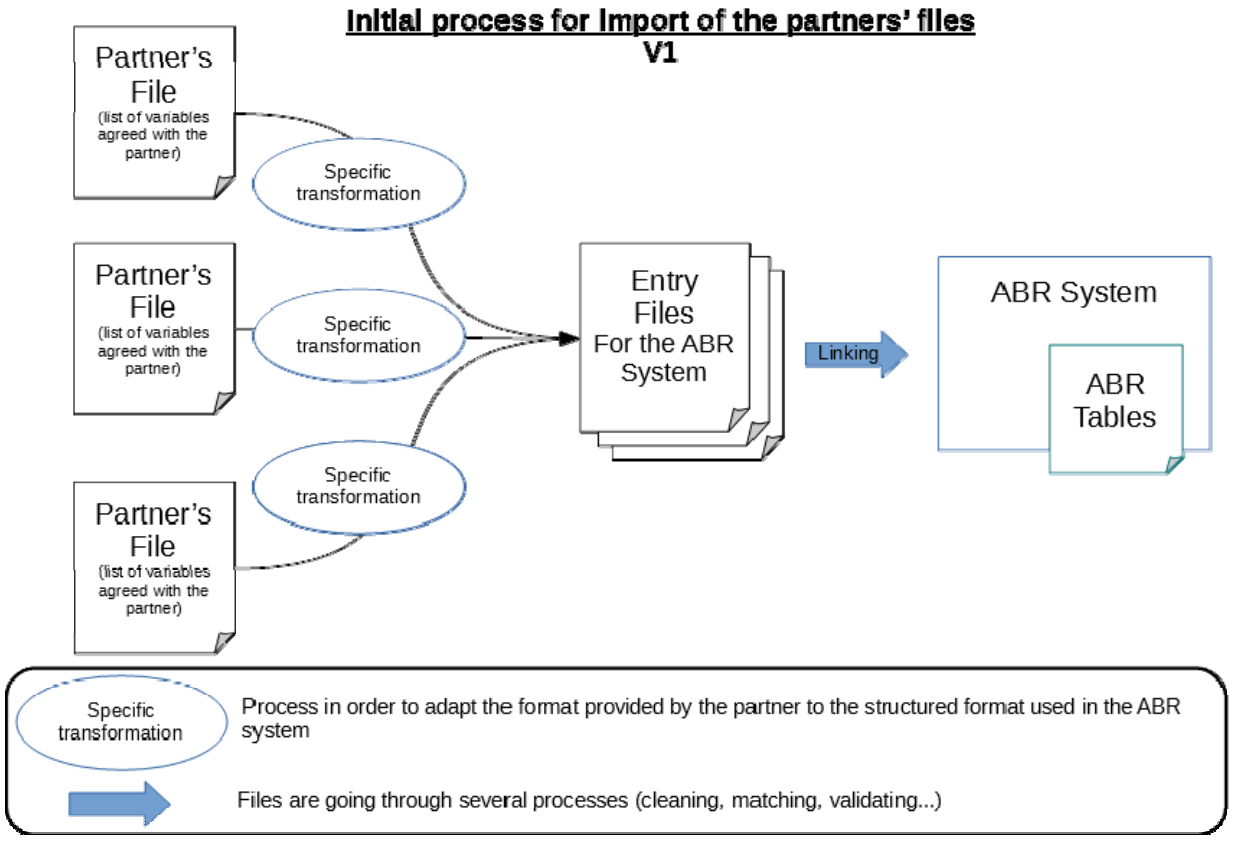
- The improvement made in data collection by each partner will be fully taken in consideration for the IT treatments concerning the choices of best values for each variable;

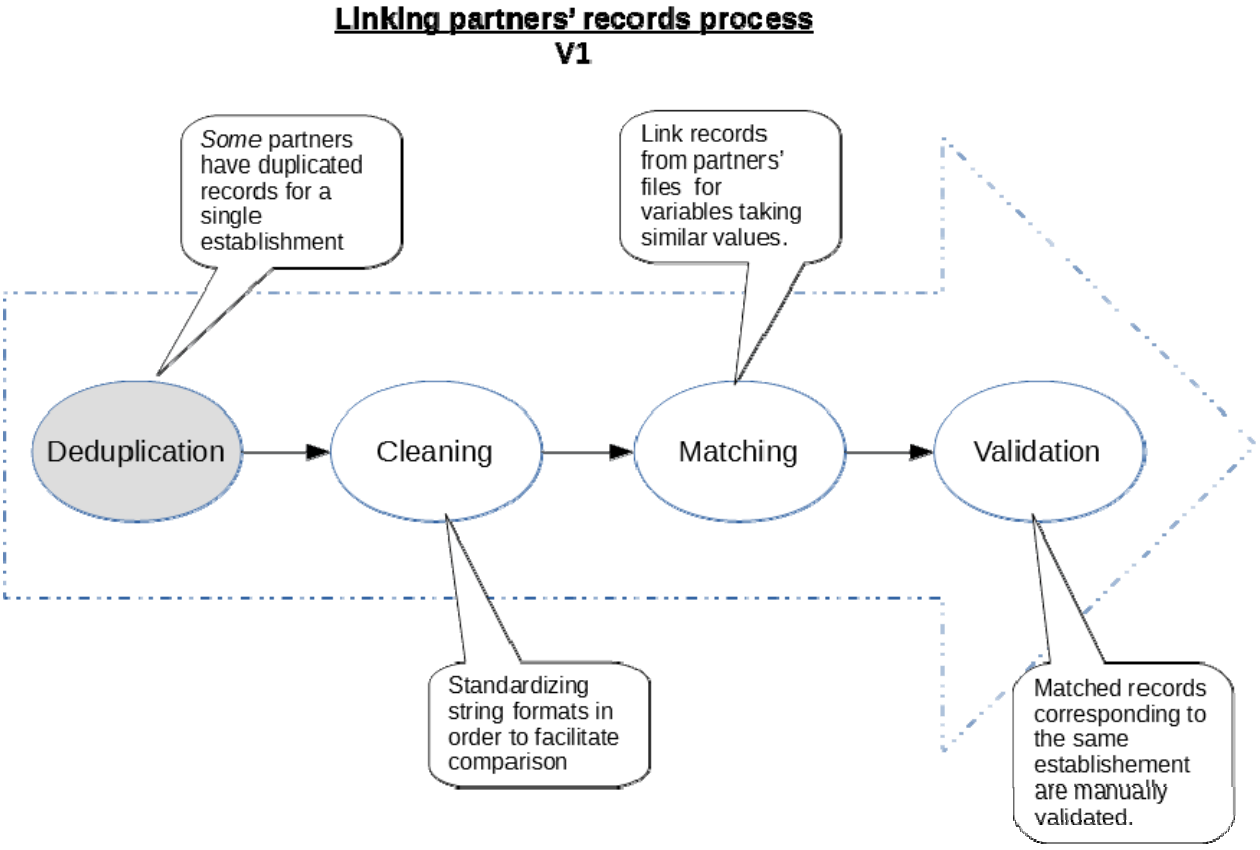
- Partners will transmit all their records with due information about the ones created since the last transmission and the already existing ones where variables have been updated.

From a technical point of view the mature system will include operational answers to:

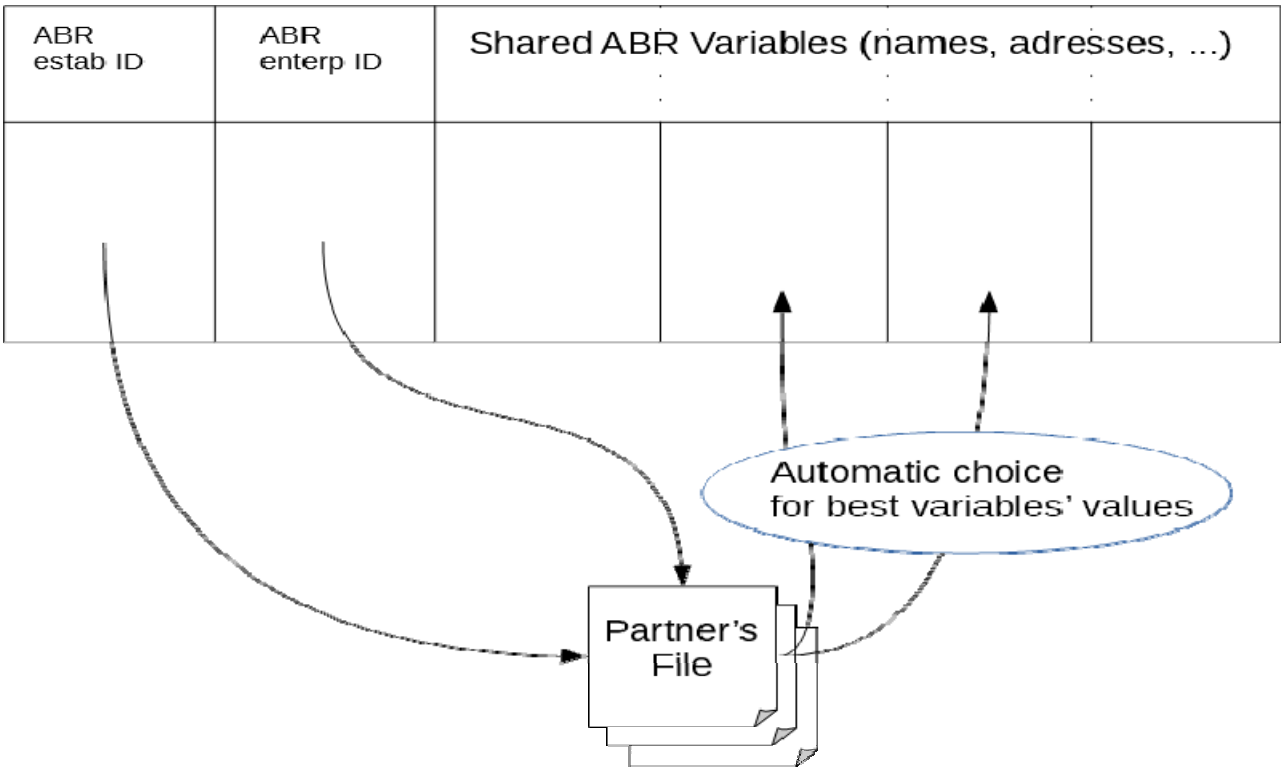
- Technology issues concerning the exchanges (way of exchanging information, best possible formats…);
- Access rights to the database for each partner;
- Ergonomic Graphical user interfaces (GUI)
- Dashboards for each partner
- Content management system (CMS)

Graphics that describe the work process (some being reproduced in specific parts)





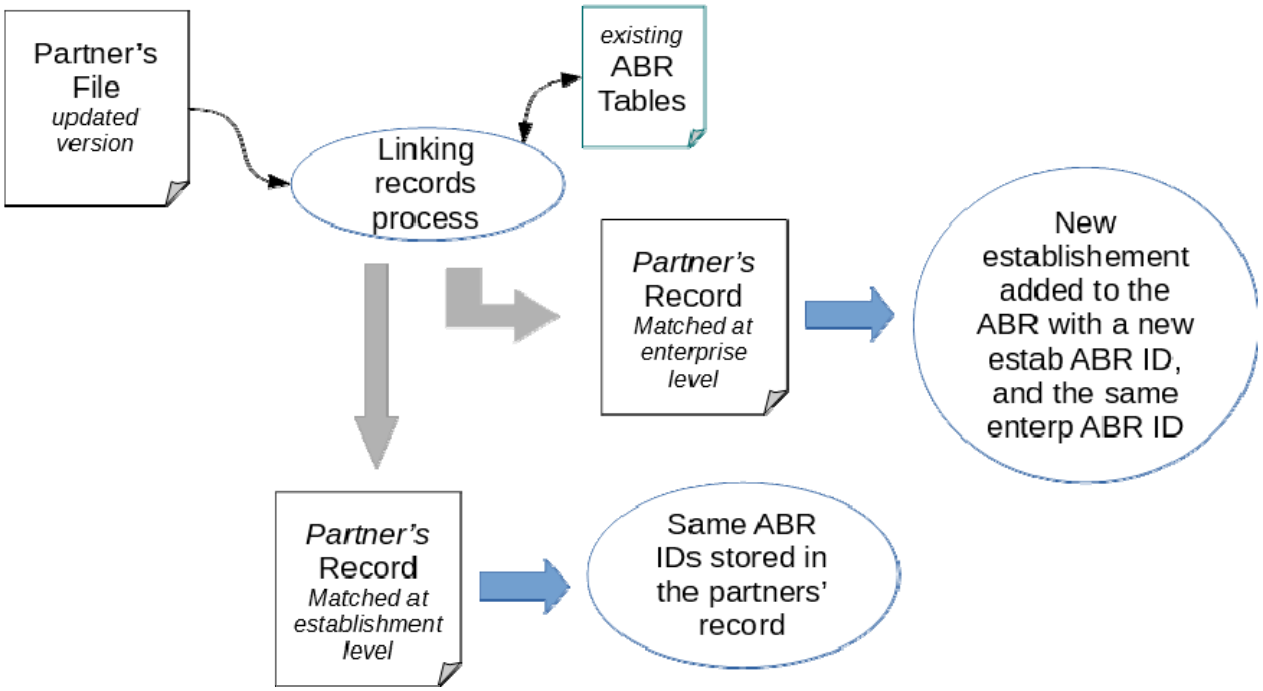
ABR Tables
V1



Update Process

Insertion of new ABR units, update of best values for variables

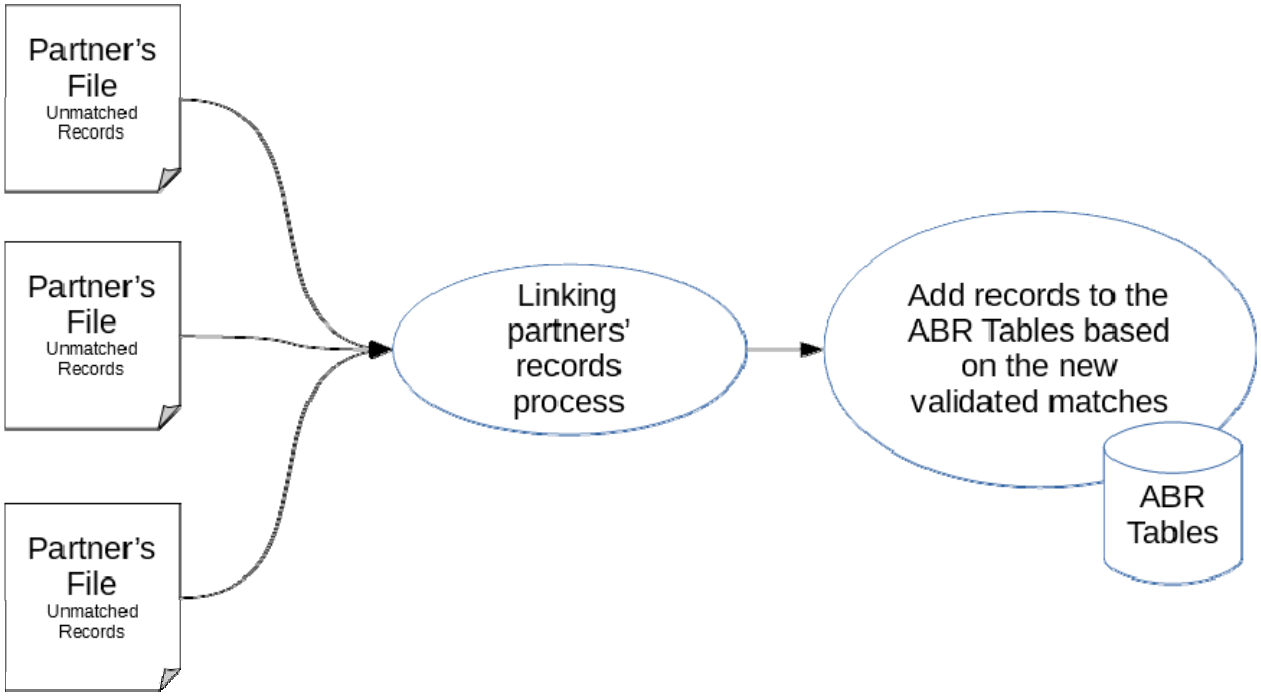
Matched Records



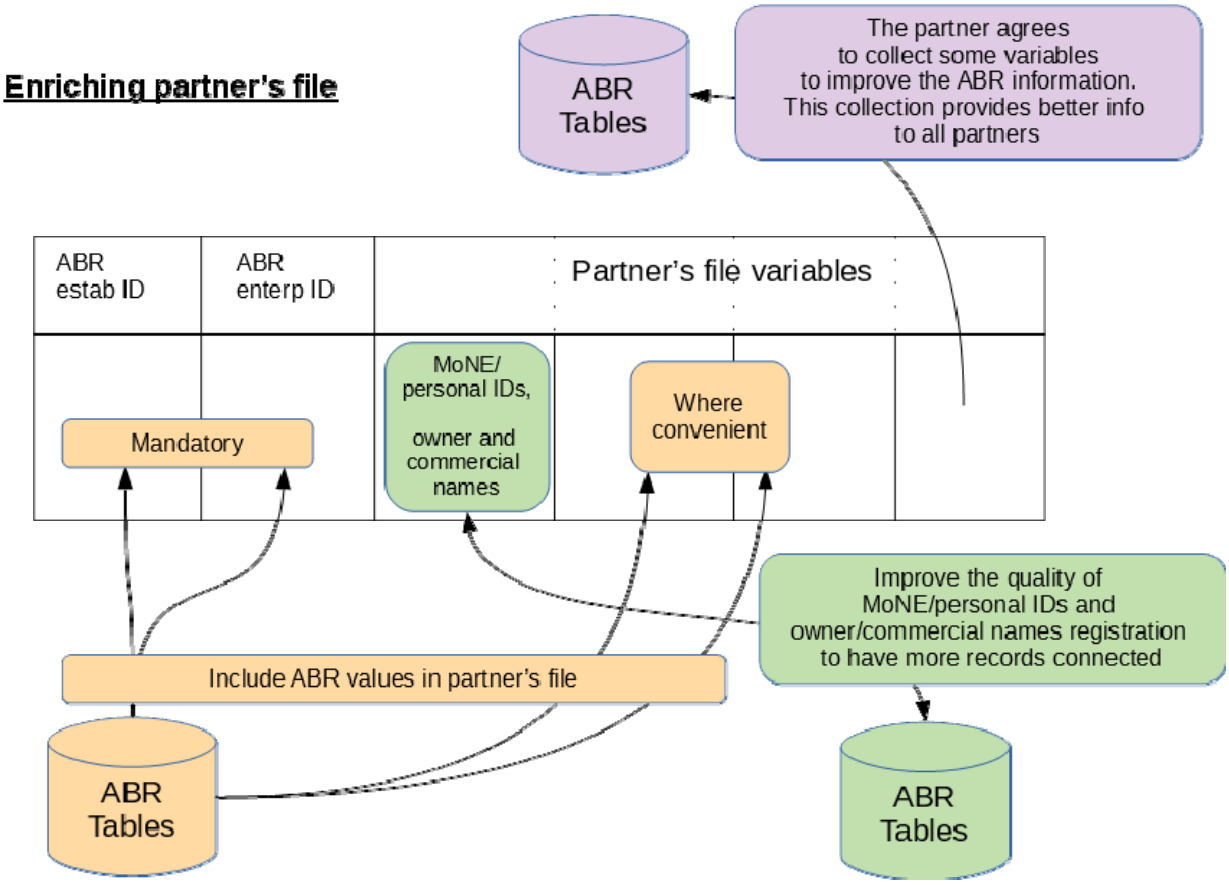
Update Process

Insertion of new ABR units, update of best values for variables

Unmatched Records



Enriching partner's file



5. The Risks and Difficulties (Challenges)

Basically, what makes the quality of the ABR is (1) the coverage of active establishments/ enterprises (including regular updating with records for the new created units), (2) the values provided by the ABR tables for the shared variables, (3) the linkage provided between ABR records and partners records, (4) the absence of duplication of enterprises in ABR records.

The risks and difficulties we are naming here under are what makes the quality at risk if one or more of the factors listed are not met.

1- The coverage of the ABR relies on the coverage of the partners' registers. The basic risk is to have no partner registering a unit that corresponds to a place of activity. As a consequence, the ABR would not register it.
   It is considered that the partner that can at best address the risk is the Municipality. It has the best ability when collecting taxes to map all the places of activity on its territory. Municipality cooperation is key to get an exhaustive coverage of active establishments in the ABR.
2- The values of shared variables rely on the information collected by each partner (what is asked, the extent of updating and its freshness).
   Some partners are more concerned than others by collecting some of the shared data. Municipalities have the best ability to collect addresses (also GIS coordinates that can be transformed in a classical address through a table). Companies ID and commercial names are the responsibility of the MoNE which has an exhaustive knowledge for those variables. Owner names, personal ID can be better known by the VAT administration; still this knowledge is not exhaustive (when people are not registered as VAT taxpayers).
   The risk is when the data collected by partners are badly registered, not updated or not exhaustive. Then the value of a shared variable in an ABR record is not known with the proper quality and, as a consequence, it is not entered in the ABR. As for its ability to collect at best the info for a variable, the given partner should be mobilized to address the issue.
3- The linkage between ABR records and partners records is key for the usefulness of the ABR.
   This issue can be addressed smoothly if efficient "matching" variables are present in each partner's file. MoNE ID, Companies names, Personal ID are the most efficient information to make secure linkages. When missing, Owner names are a useful information to use.
   The risk is that partners' records do not register properly those variables. Each partner has to commit to register them and to update records where these variables are missing in its past records.
   When created, each ABR establishment record is given two ABR IDs (ABR Establishment ID and ABR Enterprise ID) that are generated by the System. Each partner's record that has been included in a validated matching is linked to "its" ABR Establishment and ABR Enterprise records through those ABR IDs.
   If the partner does not introduce the two ABR IDs in its register for its records that have been linked to the ABR records, we face a major risk of destroying the linkages for the record at a next update phase. The partners have to be committed at introducing the ABR IDs for each of their linked record.
4- In order to have an efficient process to decide linkages between records, each partner has to provide variables it registers sometimes for its own needs. This list has to be established with each partner. The risk is that the partner does not provide enough information in order to secure the quality of the linkage decisions.
5- The absence of duplication of enterprises in ABR records is key for a proper linkage between ABR and records of partners' files. It is the responsibility of the ABR management team to ensure the proper processes to avoid duplications in the ABR tables.
6- One issue to be addressed will be the follow-up of establishments in order to update its "active" statute. This follow-up should be done by local administrations in relation with their annual fieldwork. It can be expected a high quality regarding this update in few years from now.

6. Human Resources Capacities needed to manage the ABR System

6.1 Statistical Human Resources Capacities

To manage the ABR system, the team should include at least three full time statisticians with a good capacity in data management. They will be in charge of agreeing with partners on the information to provide, to specify the process for cleaning, matching the data and validate the matching. They will also be in charge of the analyse of the quality of the data in partners' files in relation with the definition of the algorithm to choose the best available values. They are responsible for the quality of the ABR tables, the linkages with the partners' files. They have to propose the improvements that are ready to be implemented year after year.

To summarise, the statistical team will be responsible for:

- Providing specifications to the IT team on the variables to be populated on the database tables,
- As well as maintaining the content of the tables and submitting new or modified specifications as needs arise and shift ,
- Receiving and reviewing administrative data from partners, ensuring timeliness, coherence, accuracy of updates,
- Documenting concepts, processes or coverage changes over time, and producing statistical analyses of the content of the register.

6.2 IT Human Resources Capacities

To maintain and manage ABR system and database, the staff dedicated to the ABR System IT includes IT capacities in developing databases and in database management systems (DBMS), and in writing programmes. Specifically for the ABR, they shall develop .NET and Java programming and use Duke softwares.

At least one database administrator (DBA) will be required. The DBA will interact with the database management system processes and tables that make up the register, and ensure that the tables are made available under the different IT environments that shall be used. The DBA will be responsible for ensuring that the database is functional. The DBA will also need to ensure that the database can be updated and queried as required.

Given the scope and functionality of the initial ABR design, the DBA should be assisted by another IT person. Initially both will use a packaged server software (e.g., Microsoft SQL Server Management Studio).

At a later stage of development, the IT staff will also develop (or only manage if the development is outsourced) the graphical programming that will enable ABR staff to easily access and update the content of the ABR. On line accesses for partners upload and download of files will be developed in this later stage.
Funding will be need to be secured, to ensure that software licenses are maintained and that equipment — servers, PCs, networking equipment and the networks themselves — can be maintained and upgraded when needed.

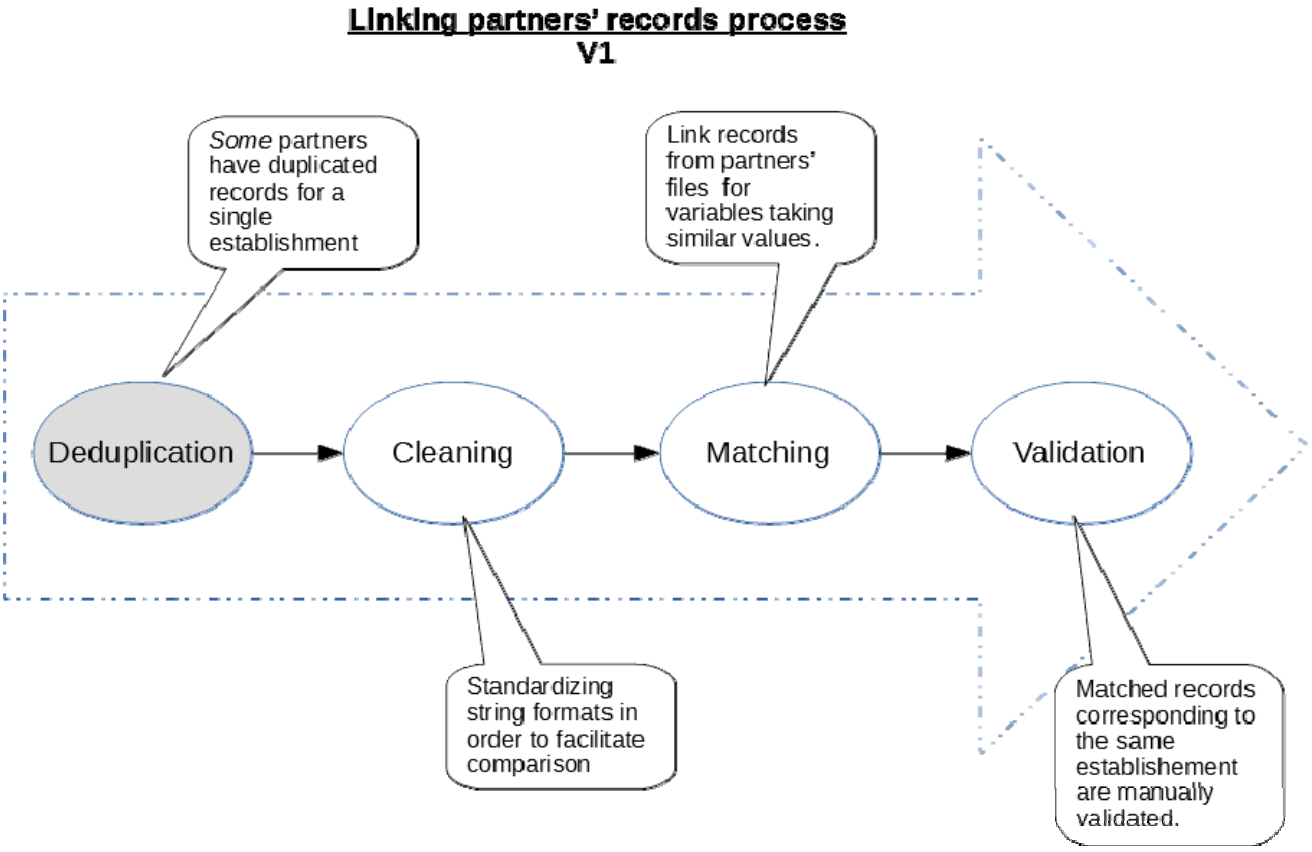
IT specialists should include

- database/data model specialists
- programming language like SQL specialists, to process large quantities of data
- .NET or JAVA specialists, Duke developers for the ABR interface (front end) and programming (back end).

7. ABR linkages management – an overview

Each partner registers in his file different variables. The quality of registration differs from one partner to the other. Also, in each partner's file there is an important heterogeneity between records for the quality of variables' value registered. Some partners have a reliable address for many records, and some not at all. Some partners have well registered personal ID when other do not.

When the ABR management team receives from a partner a new version of a file to be used for the ABR System, there are many data treatments to be realised in order the System fully benefits from it.



Type of the files: the files received may differ in formats (Excel sheets, db, ...).

As Excel format is of common use, it is desirable to ask all partners to make their data provided in this format or at least in a format easily transferable in Excel sheets.

Description the files: each file needs to be described. Each variable unit, coverage, quality, possible duplication for a same establishment/enterprise, number of values not registered,… have to be documented in order to have a proper analyse of the file's quality regarding what is important for linking its records with ABR records.

Two types of descriptions are made for each initial file or new version. A general description contains a general information about the whole file (file Name, number of records, date of transmission, date of description, number of variables, and broad coverage). A more detailed description concerns each variable (variable type, description of the variable, missing values, damaged values, classifications, range of values, duplications, frequencies, meaning of percentages in case of multiple choice questions (in a table or charts), and odd values cases).

The duplication issue: The ABR tables record each establishment and each enterprise with one record. There is no duplication.

Partners' files can have duplicated lines regarding a same establishment. For instance, in municipality files, each activity of an establishment can be recorded in a separate line. So, if an establishment has several activities, it is represented in several records. Also, with the VAT files, several taxpayers can pay for a same establishment or a same enterprise which therefore will be represented in several records.

To manage the ABR tables with no duplication problem, a de-duplication concept is implemented which consists of detecting duplicated establishment records within each partner's file thanks to the variables provided by the partner's file. The main method consists in matching the file with itself and analysing the records that contains similar values for some meaningful variables.

When returning its file to the partner with the ABR IDs for its records that matched, all duplicated lines should get these same IDs.

Cleaning the files: The data for a given variable can be registered with various formats. It can include mistakes in its registration. Before comparing the values recorded, it is needed to "clean" the data (respecting a format). For instance, regarding Owner names, we look for commas, semicolons, carriage returns, to replace them with a space (See Appendix 2).

Also, there can be mistakes in the recording which have an effect on records, i.e. Owner names in the column Commercial name. In order to compare efficiently the records, we concatenate in one variable the names (owner_name with commercial_name) to compare the information for the two variables.

Other cleaning and concatenation are needed to compare additional variables that will be used to decide if two records are truly matched (and not only "mathematically" matched). For instance phone numbers (mobile, landline, fax) can help after being "cleaned" to have the same format (with their area code).

The Matching-Linkage issue: Linking records from different files is the process to compare records based on given variables' values. For example to link municipality file' records with MoNE file's records on Commercial ID and Commercial _name. When the value is identical, the records are truly connected (a same enterprise). Nevertheless, they can differ for some other variables as addresses, which means that they are different establishments (local units).

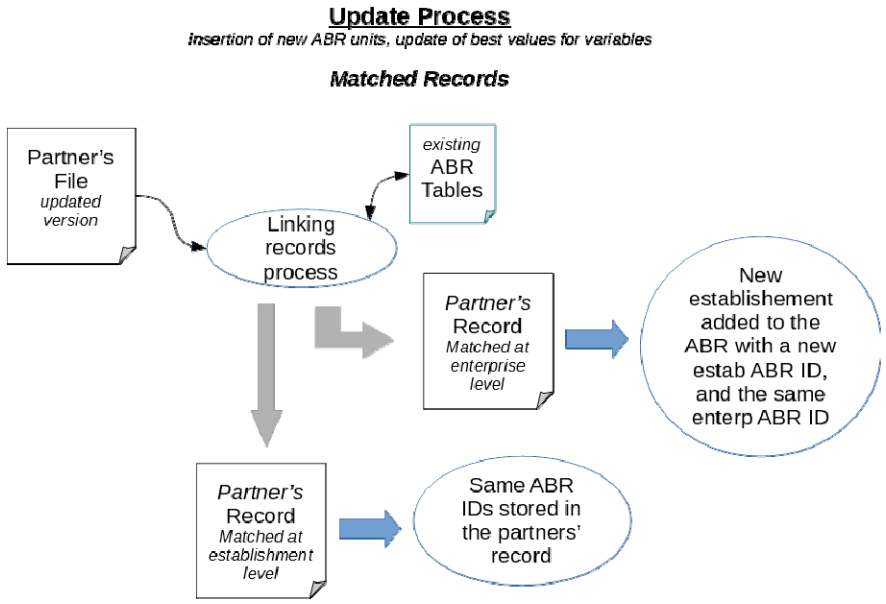
Duke software is a convenient tool to manage the matching-linkage process. It works with calculating distances between characters chains (Levenshtein distance, for example). By default, Duke uses an average of the all properties coefficient to come up with the global similarity coefficient. For our needs, we developed ad-hoc measures and comparison coefficients for calculating a distance that fits for our needs. The Duke programme outputs provide a list of records that match when their calculated distance is under a given threshold.

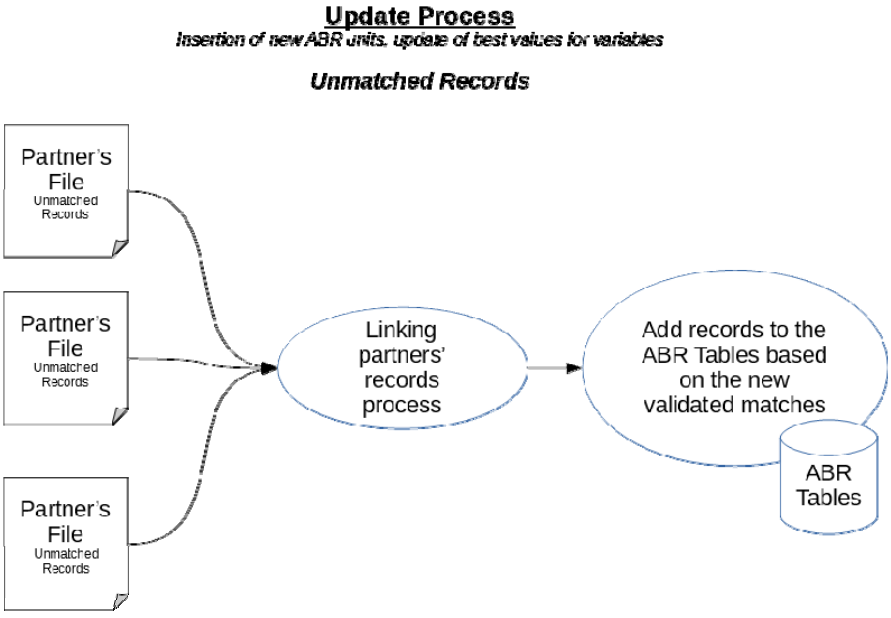
Validation process: After receiving the matching output, a team of trained validators validates the matching proposals by analysing "manually" the consistence of the values of relevant variables in the records matched. A code summarises the validation of the matching. The code created by the validator consists of two numbers (ex. 2.1). The first one indicates the records that belong to a same enterprise. The second indicates the records that correspond to a same establishment of this given enterprise.

Insertion process and partners' file records linkage: This phase consists in inserting the ABR created establishments and enterprises in the ABR tables and adding to the partners' linked records the two ABR IDs that characterise their linkage with the relevant ABR records.

This process has to be implemented with care as it includes three basic cases:

1. A partner' record corresponds to an enterprise and an establishment not yet created in the ABR tables
2. A partner' record corresponds to a new establishment of an already enterprise existing in the ABR tables
3. A partner' record corresponds to an enterprise and an establishment already created in the ABR tables, just bringing a linkage for this partner that was not yet done.





Each new partner record linked with the ABR contains variables values that can be the "best available values". A process of choosing these best values for the ABR table is run after the linkage process is done. Basic way of choosing is described in part 3.2 above. With some partners in charge of collecting fresh values regularly, the way of choosing the values will evolve accordingly.

8. Partners' roles & benefits in the ABR system

The ABR System is relying on a shared management of information made by a PCBS team (see Section 6).

The ABR System has been developed with three basic considerations:

- Each partner will keep on managing its own file with its IT system
- Each partner will keep on following its units to manage (taxpayers, companies, activities, …) which are not necessarily the two units ABR manages (establishments and enterprises).
- Establishments and enterprises are the two relevant units that allow to link all the partners' units in a rational and useful way to allow them to know if they record all the units they should be in contact with.

As a consequence of these considerations, the ABR can play its role which is to help ALL partners to select the establishments and/or enterprises whom they should be in contact with. This approach of the ABR means that there are many actual and potential users of the shared administrative business register.

The ABR System is managing information about existing economic units described by basic identification variables (official IDs – personal or company –, company or owner names, geographical location, headquarter, and will expend information on phones, year of creation). Its aim is to help each partner to know about the units it should be in contact with for achieving its own duties.

Only partners to the ABR, who provide information for its updating, have their records connected to the ABR.

The ABR System is an "internal" management system for an administrative purpose. In line with the ABR present aim, each partner gets the information about the shared ABR variables' values and gets its own data enriched by the two ABR IDs (enterprise and establishment).

When an online system will be available, each partner will be allowed to upload its data and download the ABR files and to have its own file updated with the links established to the ABR tables.

Two "types" of upload are to be done:
- the new registered units since the last update
- the new values for already existing units (in order to share updated values).

The ABR tables shall include the "key" variables that allow a direct matching between the new values for the variables and the former information: the ABR establishment and ABR enterprise IDs for already linked records; the MoNE IDs and Personal IDs for newly registered units.

The role of PCBS is to manage the ABR System with the information uploaded by partners.

PCBS is the best option for housing the ABR management team: it is a public institution that can receive the detailed information provided by the partners, guarantying a full confidentiality (acting along highest standards for the rules of information confidentiality).

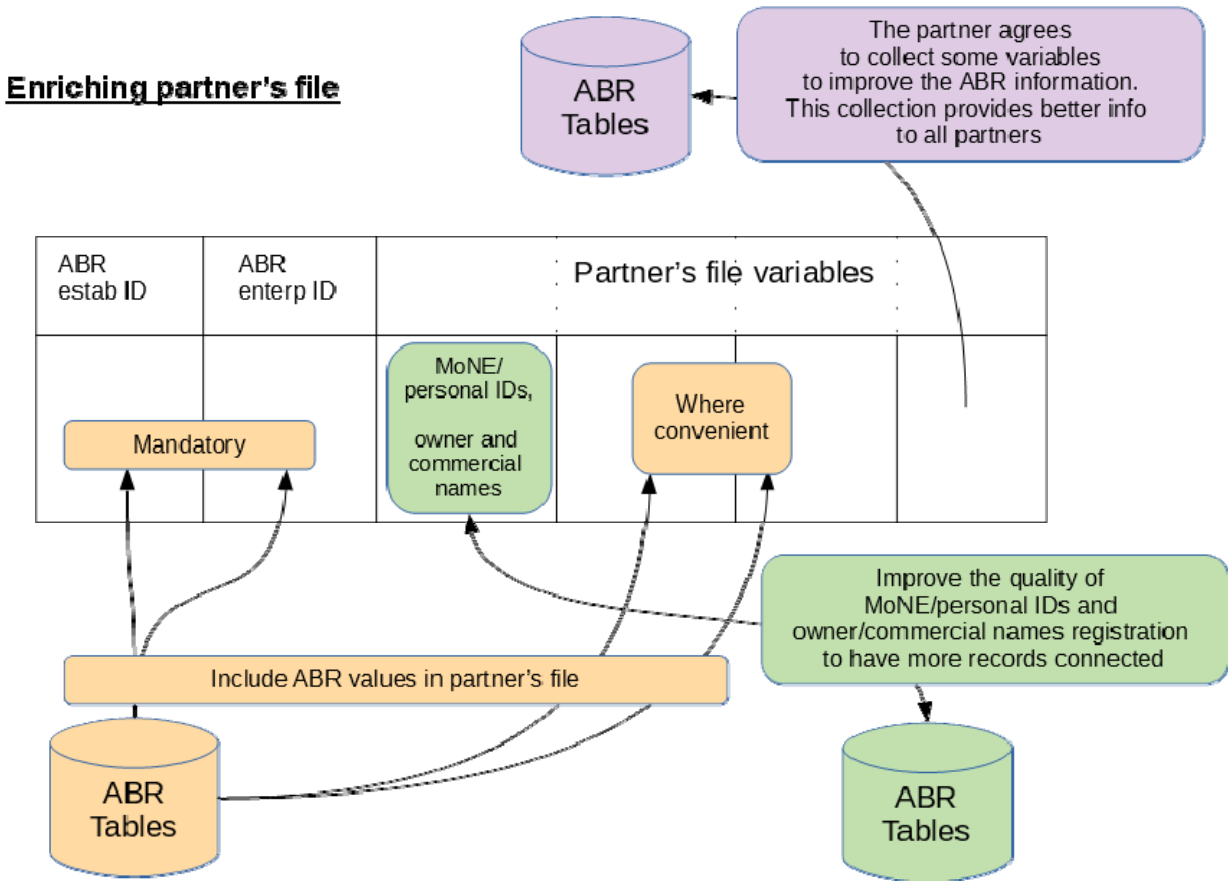
Also, at technical level, it has the needed competencies in order to manage the heterogeneous information provided by partners (the linkage of the info inside the ABR and between the ABR and each partner file, the management of complex updated information coming from various sources, …).

To do its management task, the PCBS team shall rely on the inputs coming from the partners. Each partner communicates the values for the variables he has to provide (see lists in Appendix 3). It also includes other variables needed to manage efficiently the ABR System (i.e. date of creation of the record, date of the change of a variable value in the partners' file, tax ID from the VAT file when different from the personal ID or MoNE IDs, website,…). These variables contribute to create the ABR units without creating duplications and contribute to link without mistakes their records to the ABR records.

Each partner can have in its file specific variables that other partners don't have. Some partners have a better quality for some variables they record (better updated, more details). As an example, municipalities have often a more reliable address for the establishment. Some partners do not record at all an ABR shared variable (as an example, not all partners record the personal ID).

In order to improve the quality of the shared variables' values in the ABR, some partners have a better chance to do it well and at a decent additional cost. The ABR System will benefit of their data collection if they accept some additional data collection. For instance, municipalities could be asked to collect with highest care the GIS coordinates in order to get the best info for the establishments' addresses. Also, they could be asked to put a question about the location of the headquarter if an establishment belongs to an enterprise that have several locations.

When responsible for particular variables, the institution in charge should be asked to follow the format that will be the better for matching (in order to avoid as much as possible the stage of cleaning of data).



Appendices

Appendix 1: Life Cycle of the ABR and Data Flow

Data inputs for the ABR – major challenges

The data used for the ABR are provided by the Ministry of National Economy, the Ministry of Finance and Planing (VAT), the National Chamber of commerce, ten municipalities. In the future, other partners are expected to join the providers.

Data are collected from each partner who provides his file that include a set of agreed variables. The selection is done taking in consideration what is available and what is needed to implement the ABR processes.

Each file has to be described in a standardized format that allows analysing the scope, the quality of the recording, and all characteristics related to the content of the file.

After having analysed the Palestinian situation, it has been analysed that four variables can be collected by every partner that allow the basic linkages between the records partners are registering: MoNE ID, Personal ID, Commercial Name, Owner Name.

In the reality of the files provided, the data collection of these variables is uneven. Uneven between partners, uneven within each file. A crucial challenge for the development of the ABR coverage and linkage with partners' files will result from the improvement of the data collection of these variables.

There is a large discrepancy between municipalities for recording information. The positive results of a test done with Municipality of Ramallah allow to consider that a special form could be designed for all municipalities in order to get a high quality data collection.

Many questions have to be faced with the files, such as the following:

1. Not all partners have a computerized file. The partners who don't have computerized data will not be included in the ABR system for the time being. The computerization is a prerequisite for participation.
2. Not all format in which the data of the partners are providing comply with a secure and easy way of managing them. Some widespread format should be agreed (Excel sheet or equivalent, database formats with proper description)
3. Some partners did not provide variables that are used for the matching (VAT has provided a tax ID instead of the Personal ID). Agreement with the provider has to be set in order to improve the number of partners records linked into the System.
4. Differences in **concepts and definitions** for the variables collected by the different partners have to be well described and taken in consideration in order to define a proper process for matching the records.

Data Description

A general template has been defined in order to follow routine rules for the description.

PP2 – Template for detailed documentation of files

(logo)	Detailed Study Data set name	Author Date

Data set properties

File name and format

Number of records

Unit described by a record (enterprise / establishment)

Number of variables

Transmitted by

Transmission date

Number of duplicates, and definition of how duplication is defined (as a variable or list of variables) ; number of occurrences for each duplicate.

Other reference documents (other description, web site, form for data entry, etc.)

Variables

(For each variable, in the order of the file)

Variable name (Arabic and English) : as in file

Normalized variable name (English,)

Meaning of variable

(Short description)

Type of variable (numeric, string, date)

Complement on type (e.g. free text versus classification labels or code, discrete or continuous variable, conformance to a syntax or regular expression, etc.)

If variable is code, description of the code list: name, version, maintaining agency, level of standardization.

Number of missing values (and how missing values are represented).

Description of irregular cases with examples for each case.

Simple descriptive statistics (min, max, frequency table...)

Other comments on the data quality.

Links with other variables.

Suggested cleaning procedure specification.

Each new file provided (update of a former file of a same partner or getting a file from a new partner) shall be described with all its effective features.

The description includes the elements to judge the quality of the partner files for what regards the processes to be implemented in the ABR System.

It encompasses the processes from “cleaning” the data to validating the linkages suggested by the programming.

It also includes the decision making for the automatic choice of the best values available that are registered in the ABR (from which partner can we take the governorate, the locality, a telephone number, etc.). Appendix 3 gives more details about how ABR-V1 has selected the values for variables from the partners’ files

Cleaning the Data

The “cleaning” process consists of harmonising the presentation of the values taken by a variable in order to get a better accuracy of matching similar values taken in the different files.
For instance, cleaning a phone number is proceeded by omitting all the signs and letters that appear in its registration and putting the area code if absent. Cleaning the Commercial names is very much needed as they are often entered with mistakes. Some specifications can be programmed in order to keep only “significant” chains of characters.

A set of rules are conducted to be applied on different files from different parties to make the data presentation more homogeneous all through the records of all partners. It is particular relevant to measure similarities between the chains as there is much flexibility of using different letters for a same name in Arabic language. (See Appendix 2: Cleaning Specification)

All the cleaning specifications are implemented in a cleaning program using Java so the process is done automatically before the matching phase. This cleaning phase allows to get a better matching result for telephone, owner name, commercial name.

Data Matching

After data description and cleaning, an automatic matching is launched in order to get proposals of similar records in the various partners’ files.

The matching process is based on comparing the “matching variables” values present in the various partners’ files. The distance between two values is given a “score”. When the distance is measured under a given threshold the linkage is established in order to be validated at the next stage. When the distance is too large, records are deemed considered as different.

In order to compare the values taken by records, it has been built an integrated program using "**Duke**" (inside Java) which allows to match files even there are more than two.

The programme’s thresholds are to be adapted in order not to miss “true matches” (the threshold’s value is not to be set too high), nor to get too many links proposed that are “false matches” after the “manual” validation (the threshold’s value is not set too low). To say it another way, we face erroneous matches when the matching rules are ‘too loose’ and a too low number of matches when the rules are ‘too rigid’.

Manual Validation

After getting a file of proposed matching as a result of running the Duke program, the matchings have to be validated manually by a trained team.

Several variables are used to take the decision such as: ID, commercial name, owner name, phone numbers, activity, and the address.

The validation code consists of two “numbers” in the form " X.Y". X is given to each record belonging to a same enterprise and Y is given for each record that belongs to a same establishment.

All the establishments which belong to a same enterprise will be grouped together by having the same X regardless the location of the establishment.

Branches (establishments, local units) are defined at this stage by checking their particular address. The first address is noted 1, the second 2, etc.

Attribution of two ABR IDs

The validation codes are translated into ABR IDs, one for the enterprise (E + 8 digits) and one for the establishment (L + 8 digits) with are introduced in the ABR system (the ABR tables and the partners’ records that are linked to this ABR tables’ record).

Choosing the best Values for the variables

Depending on the characteristics of each partners’ files, and the recording or no recording of a value, there is a programme that choose what can be considered in an automatic treatment as the best value for each variable.

For instance, the values recorded in MoNE file are always taken as the best values for the MoNE ID and the Commercial name (as this file is exhaustive regarding companies and is delivering the Commercial name.

For addresses, Municipalities’ files are considered as a first choice, VAT file come after.

(See Appendix 3) for more details.

Appendix 2: Cleaning Specifications

Specifications were designed/defined to have more homogeneous chains of characters in order to compare values between records. The simplification that the specifications do are not altering what is significant in the values recorded by partners. The homogeneity is increasing the probability of matching, especially of true matches.

Step1

Replace some letters and signs by other ones in commercial name and the owner name as the following table:

Actual case	Replaced by	before replacement	after replacement
- - _ * / \ = +) (& ^ % $ # @ ! ~ " ' : ; ؟ ?> < [ ] ، . , ’ } { ْ ٍ ِ	space	" مخمر موز "	مخمر موز
& between two Arabic words	و	كيك & بيك	كيك و بيك
& between two English words	and	art & glft	art and glft
ة	ه	كافتيريا الامانه	كافتيريا الامانة
آ- إ -أ	ا	تكسي المأمون	تكسي المامون
لأ-لآ-لإ	لا	شركه الإعتماد	شركه الاعتماد
شراكة	space	شراكة الإعتماد	الإعتماد
شركه	space	شركه الاتصالات	الاتصالات
شركة	space	شركة نيوستايل	شركة نيوستايل
شراكه	space	شراكه سبيس تون	سبيس تون
الشركة	space	الشركة الفلسطينية للأثاث	الفلسطينية للأثاث
الشركه	space	الشركه العربية للطباعة	العربية للطباعة
واخوانه	space	الأشمر واخوانه	الأشمر
وإخوانه	space	ابومحمد وإخوانه	ابومحمد
واولاده	space	صبحي الحداد واولاده	صبحي الحداد
وأولاده	space	صبحي الحداد وأولاده	صبحي الحداد
محلات	space	محلات الأخوة	الأخوة
محل	space	محل السعادة	السعادة
4 in Arabic word	فور	4 يور هوم	فور يور هوم
4 in English word	For	4 me	For me
م.خ.م	space	شركة الاسلاك المعدنية م.خ.م	شركة الاسلاك المعدنية
م. خ.م	space	شركة الافق للتكنولوجيا م. خ.م	شركة الافق للتكنولوجيا
م.خ	space	شركة مطبعة بيت المقدس للطباعة الحديثة م.خ	شركة مطبعة بيت المقدس للطباعة الحديثة

شركة المجموعة الاهلية للتأمين	شركة المجموعة الاهلية للتأمين م.ع.ع	space	م .ع.ع
شركة التكافل الفلسطينية للتأمين	شركة التكافل الفلسطينية للتأمين م.ع.م	space	م.ع.م
شركة بنك فلسطين	شركة بنك فلسطين م ع م	space	م ع م
شركة م ع كواملة للتجارة والاستثمار	شركة م ع كواملة للتجارة والاستثمار	space	م ع
شركة احمد ابو عين للتجاره و الاستثمار	شركة احمد ابو عين و شركاه للتجاره و الاستثمار	space	و شركاه
كفيتريا الامانة/ 2002	كفيتريا الامانة/مغلق ابو علي 2002	space	مغلق ابو علي
يافا للحوم المجمدة/ 2002 - ---------	يافا للحوم المجمدة/مغلق سفيان 2002 ----------	space	مغلق سفيان
مكتبة العودة للسياحة/ 2004/7	مكتبة العودة للسياحة/مغلق سلطان 2004/7	space	مغلق سلطان
مؤسسة ماك للتجارة العامة /	مؤسسة ماك للتجارة العامة /مغلق	space	مغلق
ابو عيشه للهندسه والمقاولات/2003/8	ابو عيشه للهندسه والمقاولات/غير موجود ابو علي 2003/8	space	غير موجود ابو علي
عيادة خاصة/ 2002	عيادة خاصة/غير موجود سفيان 2002	space	غيرموجود سفيان
*للاتصالات MICRO NET prime	*للاتصالات MICRO NET prime غير موجود مكانه	space	غير موجود مكانه
جوال سنتر/ 2003/9	جوال سنتر/لا يوجد ابو علي 2003/9	space	لا يوجد ابو علي
نقليات الحسنين / 2002	نقليات الحسنين /لا يوجد سفيان 2002	space	لا يوجد سفيان

Step2

Connect some words with the next word by removing the space which exists between the two words to be one word as the following:

The word	Before connecting	After connecting
ابو	ابو محمد	ابومحمد
أبو	أبو أحمد	أبوأحمد
عبد	عبد الكريم	عبدالكريم
دير	دير ابوضعيف	ديرابوضعيف
بير	بير زيت	بيرزيت
كفر	كفر عين	كفرعين
عين	عين يبرود	عينيبرود
نيو	نيو ستايل	نيوستايل
بيت	بيت لقيا	بيتلقيا

Step3

There is some couple of words that must be aggregated in one word without space between them. The space between them must be removed as the following:

The first word	The second word	Before combining	After combining
سوبر	ماركت	سوبر ماركت	سوبرماركت
ميني	ماركت	ميني ماركت	مينيماركت
لتاجير	السيارات	لتاجير السيارات	لتاجيرالسيارات
لتأجير	السيارت	لتأجير السيارات	لتأجير السيارات
New	look	new look	newlook
NO	name	NO name	Noname
Play	station	play station	playstation
Sea	sons	sea sons	seasons
Too	u	too u	toou
Book	stores	book stores	bookstores
كمبيو	نت	كمبيو نت	كمبيونت
اند	شوب	اند شوب	اندشوب
بانور	اما	بانور اما	بانوراما
بلي	ستيشن	بلي ستيشن	بليستيشن
جراند	فارم	جراند فارم	جراندفارم
كوزمو	بولي	كوزمو بولي	كوزموبولي
جلو	بال	جلو بال	جلوبال
سبيس	تون	سبيس تون	سبيستون
ستا	لايت	ستا لايت	ستالايت
سي	جي	سي جي	سيجي
تكنو	بال	تكنو بال	تكنوبال
سي	دي	سي دي	سيدي
انتر	بال	انتر بال	انتربال
انتر	ناشونال	انتر ناشونال	انترناشونال
اي	فون	اي فون	ايفون
او	تو	او تو	اوتو
اوفر	سيز	اوفر سيز	اوفرسيز
يا	هلا	يا هلا	ياهلا
بلو	سكاي	بلو سكاي	بلوسكاي
بي	تي سي	بي تي سي	بيتيسي
فن	تاستك	فن تاستك	فنتاستك
فور	يو	فور يو	فوريو
دراي	كلين	دراي كلين	درايكلين
فودا	فون	فودا فون	فودافون

Step4

After splitting the words, search words with only one letter and make one word as follows:

The sentence	The first word	The second word	The result
مكتب السياحه و السفر	و	السفر	مكتب السياحه والسفر

Step5

Replace Capital Letters by small Letters for English words

Before connecting	After connecting
BLUE DRY CLEAN	blue dry clean

Appendix 3: Variables

A. ABR variables to be shared with partners

The **ABR variables that are shared with all partners** are public information.

The list of variables that can be shared with partners to improve the quality of their data and files are the following:

- ***Commercial Name for companies.***
- ***MONE ID: for companies.***
- ***Owner Name for non-companies.***
- ***Personal ID for non-companies.***
- ***Governorate.***
- ***Locality.***
- ***Street.***
- ***GIS Coordinates.***
- ***Headquarter (*).***
- ***Year of creation.***

(*) For multi-location enterprises, a headquarter information is proposed corresponding to one of the public addresses. This "headquarter" information present in the ABR table has no legal status.

The ABR system includes also two other variables that are generated by the System: the ABR establishment ID and the ABR enterprise ID. These IDs are shared with all partners. They are to be added to their linked records by partners in their own registers for allowing proper management of the ABR update operations.

Each partner records in its register variables that are identical (or similar) to the ABR shared variables. It also manages many other variables that are specific to its management needs. Most of these variables are covered by confidentiality rules.

These additional variables are not to be shared, with no exception, with the other partners to the ABR System (the ABR System secures that these variables values provided are protected from any dissemination to other partners). But it is crucial that they are provided to the ABR system as they facilitate the matching and validation and therefore ensure a higher quality matching.

Only the shared variables values, because they are all of public access, are shared in the ABR tables

B. Partner's variables

Here under are presented the variables' tables of three crucial partners for the ABR (MoNE, MoFP-VAT, and Ramallah municipality – this municipality serving as a model for what can be expected from any municipality). The tables present provided variables (in blue) and variables that the ABR system would benefit from if the institution was starting collecting them (in rose).

Four variables not collected to date by any partner should be made available to the ABR system: number of employees, date of creation of the establishment, GIS coordinates, and headquarter. The number of employees informs about the size of the unit, the GIS coordinates inform precisely on the location (they can be transformed in addresses usual form using correspondence tables), date of creation informs about newly created establishments (when this date is not available, a second best is the "date of registration" for each newly created unit's record), and headquarter informs about the location of the general managing of the enterprise.

Mone		
Type of the variables	**Variables**	**Use for the ABR**
Provided variables (registered at the moment of declaration, not updated)	Commercial_ID	MoNE value
	Owner_ID	
	Commercial_name	MoNE value
	Owner_name	
	Phone	
	Mobile	
	Fax	
	Mail adress	
	Governorate	
	City	
	Locality	
	Street	
	Building	
	Postal box	
	Activity	
	Legal_status	
Expected variables (to be provided by MoNE)		

VAT (for after V1)		
Type of the variables	**Variables_ABR**	**Use for the ABR**
Provided variables (registered at the moment of declaration, not updated)	Commercial_ID	
	Owner_ID	VAT value first
	Commercial_name	
	Owner_name	VAT value first
	Phone	
	Mobile	
	Fax	
	Website	
	Mail adress	
	Locality	VAT value second
	Quarter	VAT value second
	Street	VAT value second
	Building	VAT value second
	Floor	VAT value second
	Unit No.	
	Postal box	
	Activity	
Expected variables (to be provided by VAT)	Governorate	
	Date of creation	
Municipality_Ramallah (for after V1)		
Type of the variables	**Variables_ABR**	**Use for the ABR**
Provided variables (registered at the moment of declaration, not updated)	Commercial_ID	
	Owner_ID	Municip value first after VAT
	Commercial_name	
	Owner_name	Municip value first after VAT
	Phone	Municip value first
	Mobile	Municip value first
	Governorate	Municip value first
	Locality	Municip value first
	Street	Municip value first
	Building	Municip value first
	Unit No.	Municip value first
	Activity	Municip value first
Expected variables (to be provided by Municip)	Number of employees	Municip value first
	Date of creation	
	GIS coordinates	Municip value first
	Headquarter	Municip value first

Main characteristics of the partners' files and their variables (summary):

Each partners' file has characteristics that are taken in consideration by the programming of the ABR System in order to make the proper treatments to improve the matching process and to choose the best available values for the ABR tables' variables.

The following table summarizes the main relevant characteristics of the files:

Partner	Comments
Ministry of National economy (MoNE)	▪ There is no duplication in the register. ID is unique for each company (legal unit). ▪ There are no branches (the register is based on enterprises, not on establishments). ▪ The address registered has not to be the effective place of the activity and it happens often to be different from any effective location
Chambers of Commerce (CC)	▪ There is no duplication in the national register. ▪ In certain governorates is registered not the enterprise but an important local branch.
Municipalities (MUNIC)	▪ Municipalities' files register activity for the establishments (local units) active on their territory. So there can be duplication of establishments in several records for the several activities pursued. ▪ The addresses when registered are reliable. ▪ All the branches located on the Municipality territory are inserted in its file (field work is conducted systematically). ▪ Linkage between establishments of a same enterprise, located in different municipalities is not provided. ABR tables are a value added for that. ▪ Personal IDs in the municipalities files, when registered, are reliable
VAT	▪ The registration unit is the taxpayer. There is one record for each VAT collected. ▪ It registers some units (taxpayers) with physical address. ▪ It registers some units (taxpayers) with no physical address. These units are not registered in the ABR. ▪ It is optional for the units (taxpayers) to register several activity locations separately or under one record. ▪ To fully benefit from VAT information, it is needed to get from VAT the personal ID or MoNE ID for each record. The Tax ID (combination of personal, MoNE and Internal-to-Vat IDs) is not sufficient.

Appendix 4: The developments made for the ABR-V1 that shall be integrated in the ABR System next phase

When building the ABR System, for each new version, we start from the existing version that integrate the algorithms (and the instructions for manual validation) that have been developed for the cleaning, matching/linking, validation and insertion steps.

From an IT point of view, they are additional issues to be addressed. Beyond including full features of cleaning, matching/linking, validation and insertion steps, the issue of the current format some partners used in order to provide their data for the V1 should evolve in a way that secures the integration of their data within the ABR System. This is particularly relevant for MoFP-VAT file.

Basic ABR-V1 features are described with below.

Database

Building the database tables of ABR system V1 and the relations with partners' data, we used SQL language to build the database.

(see Appendix 7 for more details about the tables included in the existing database, relations, partners tables, ABR tables)

ABR-V1 database consists of tables that are described here under:

- **Input tables**: these are the tables filled with data received from the partners. Each received partner file is stored within a table created for the purpose. The records from the partners tables will feed the ABR tables with the best variables' values according to the defined rules.

- **Validation table**: the validation tables hold the blocks resulting from the matching process and contain some additional variables which are fed during the manual validation process. The ABR creation process will rely on those variables to gather adequate information from different sources.

- **Audit table**: the main purpose of this table is to store any modification (insertion, deletion or update) affecting the ABR tables by giving information of the modification date, the user who made the change and the values that have been changed.

- **Variables' source table**: this table holds information about the source of each variable value given in the ABR tables. Actually, when inserting a new record within the ABR table, the information related to the source of each values will be held by this table.

- **Assistant tables**: to help manage the system, the groups of users and users' permissions.

Database Tables

- **Sheet1:** this table is the main table for the Validation system and it includes all linkages validated after the matching. Each record corresponds to a record of the partner file when there has been a linkage validated for this partner. The attached variables are the validation code, the block ID which corresponds to the block identifier according to the matching program's output (Duke) and the ABR IDs generated.

  The variables managed in this table are: ABR_IDs, MoNE_ID_number, Personal_ID_number, commercial_name, owner_name, all the address variables, activity, telephone, economic_organisation, establishment_ownership, number of employees, status, activity year, legal status, registration date.

- **Establishments table:** this table contains all the establishments that are generated with the links of Sheet1.

  Each record contains: ABR IDs generated, the best values selected from the linked partners' files records for the shared variables of the ABR System (along a methodology developed here-under).

  Each establishment is unique in this table and should never have a duplicate created.

- **Enterprises table:** this table contains all the enterprises related to the establishments that are generated within the Establishments' table.

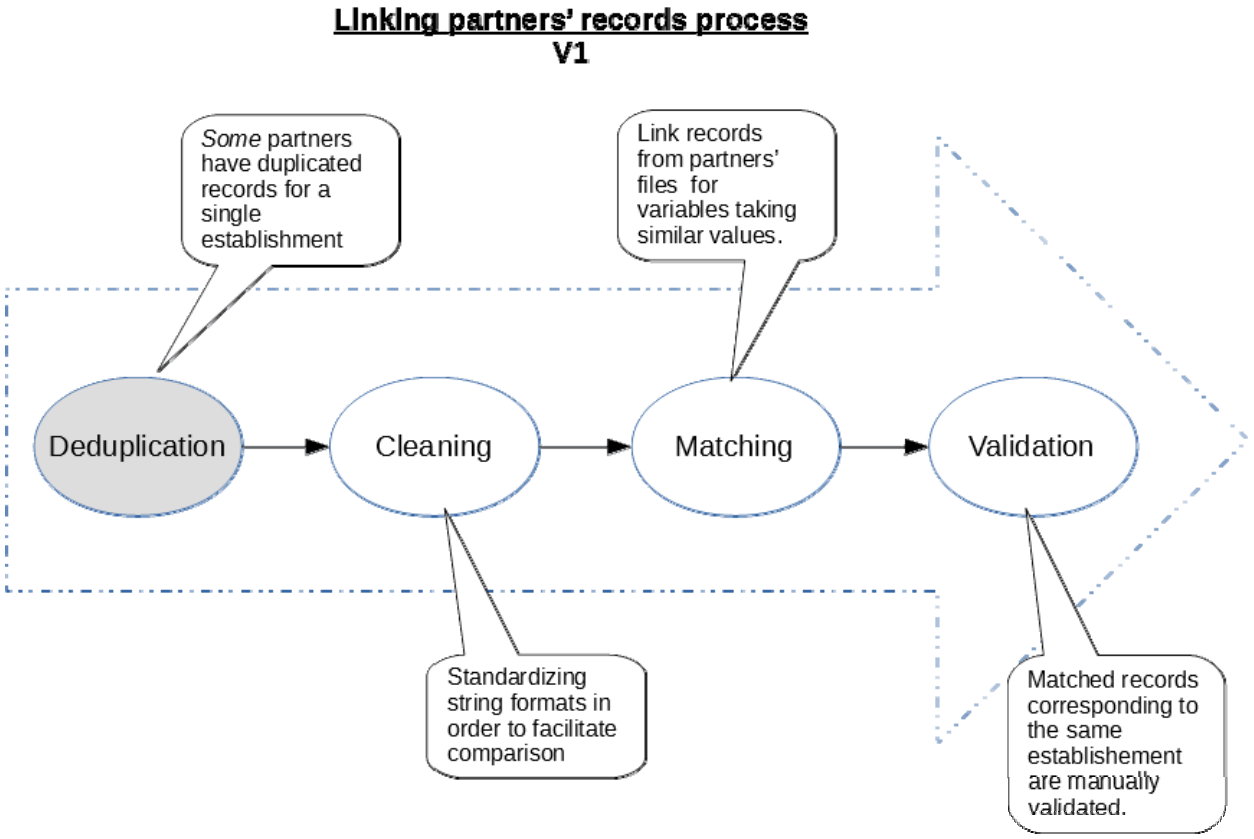
  This table contains only the ABR Enterprise ID and the ABR establishment ID for the establishment that has been identified as the headquarter during the validation process. Otherwise the ABR establishment ID variable is kept empty.

  An enterprise can be repeated in this table as many times as it has branches.

Partners tables

Each partner has one table that includes all records provided and all the variables the partner has attached to each record. The variables are enriched by the ABR_ESTABLISHMENT_ID and ABR_ENTERPRISE_ID that have been generated by the system for the records that are linked (true match).

Records from partners' files that have been temporary removed for the matching phase due to the fact that they are just a duplication of a record taken in consideration for the matching process have also to be enriched with the ABR_ESTABLISMENT_ID and ABR_ENTERPRISE_ID that corresponds.



Cleaners, comparators, deduplication matching and insertion algorithms

We build cleaning and matching algorithms using Duke to clean and match the files' records that have to be linked (two or more files).

A ***cleaner***'s job is to make comparison more relevant by removing from data values variations that are not likely to indicate genuine differences. For example, a cleaner might strip everything except digits from a zip code. Or, it might normalize and lowercase addresses. Or, translate registered dates into a unique date format. (See Appendix 2)

A cleaner can also be a combination of several other cleaners. In our case, we use the "com-own-cleaner" to clean the commercial name and the owners' name values. This cleaner is composed of other cleaners such as "TokenCleaner" and "ArabicTextCleaner". For example, the "ArabicTextCleaner" delete some terms like "ال" or "و" from the original text or normalizes certain characters such replacing all "ة" by "ه" or ("ت", "ي ", "ث") by "ب". The "TokenCleaner" does the same but instead it takes the terms list from the XML file.

A ***comparator*** can compare two string values and produce a similarity measure between 0.0 (meaning completely different) and 1.0 (meaning exactly equal). As a same information can be represented by string values that differ slightly, we have to compare beyond a "fully equal"-"different" judgment. A proper level of similarity has to be taken in consideration in order to validate a true match.

Also, different kinds of values (ID numbers, addresses, …) must be compared differently. And, comparison of complex strings like names and addresses is a whole discipline in itself.

Several comparators are provided within Duke, but we developed specific comparators for our use regarding commercial and owner names, and telephone numbers.

The automatic phase selects possible true matches. A manual validation is always necessary to finalise the decision.

Deduplication is the process to detect duplicated records regarding a same establishment within a single file. Duplication can happen for different reasons. One is that the unit followed by the partner is different from establishment or enterprise. For example, for an establishment that performs several activities, a partner can record as many lines as the number of activities of that establishment.

The usual process consists in linking records of a same partner (linking a file with itself, for example Municipality-Ramallah-2018 with Municipality-Ramallah-2018). The matching is done by comparing the values of relevant variables. A set of variables with identical values indicates the duplicated establishments.

In the deduplication process we will select one record for a given establishment which will be used in the linkage process.

But, at the end of the validation process and creation of the two generated ABR establishment and enterprise IDs, we will link all the partners "duplicated" records in the partner's file.

Matching is the process to compare two different files' records based on the values of a given set of variables. For example, we match municipality file' records with MoNE file's records on MoNE_ID and commercial_owner_names.

After calculating a distance (Levenshtein, for example), Duke provides a list of possible true linkages that have now to be considered in a manual "validation" step in order to confirm or deny the true match.

Validation system

The **validation process** consists of deciding if two records belonging to two files correspond to a same establishment, a same enterprise whatever is the unit the partner records (taxpayers, establishments' activities, companies, …).

To achieve the validation step, an IT programme has been developed that has to be "answered" with codes in order to allow the further steps to be implemented through next IT processes.

The **validation System** is a web-based system designed using ASP.net. Its main objective (give the needed info to code the different proposed linkages) presents in a user-friendly way the blocks of records that were matched with Duke. It allows to introduce codes that indicate for each matched record if it represents the same establishment than another record from another file or if it is a different establishment of the same enterprise, or if it is a false match.

These codes are introduced by a trained validation team. All matched records that are analysed by the validation team as not belonging to a same enterprise are coded as "false match".

The validation system takes the data **block by block** (a block aggregates all records from all compared files that the Duke programme has matched with a record of another partner in a direct way or through another linkage record (indirect way)). (See Appendix 6 to read more about validation system)

The validation process consists basically to look (manually) at all the **matched records** in the same block and based on their variables' values in order to decide which are **true** and which are **false**.

When a validator detects a true match between two records, it can be a similar establishment or different establishments for a same enterprise. A code that consists of two parts is given to each of the true matched records. The first part gives a same number to all the records matched because they correspond to a same enterprise. The second part gives the same number for records that correspond to a same establishment. Each establishment has a code different from the one representing another establishment.

For example, a code "201.04" means that the record is coded 201 for its enterprise belonging and 04 for its establishment belonging in this enterprise.

After finishing a block's validation, the validator submits the block to save the codes she/he entered in the database and that automatically generates a unique **ABR_ESTABLISHMNET_ID** for each establishment in the establishment table and generates an unique **ABR_ENTERPRISE_ID**. The structure of these codes are a letter and eight digits (the last one is a control number for the seven first). The letter characterises the Enterprise (E) level and the Establishment (L) level. The seven first ABR-IDs digits for both enterprise and establishment levels are not meaningful.

Insertion process and partners' files update.

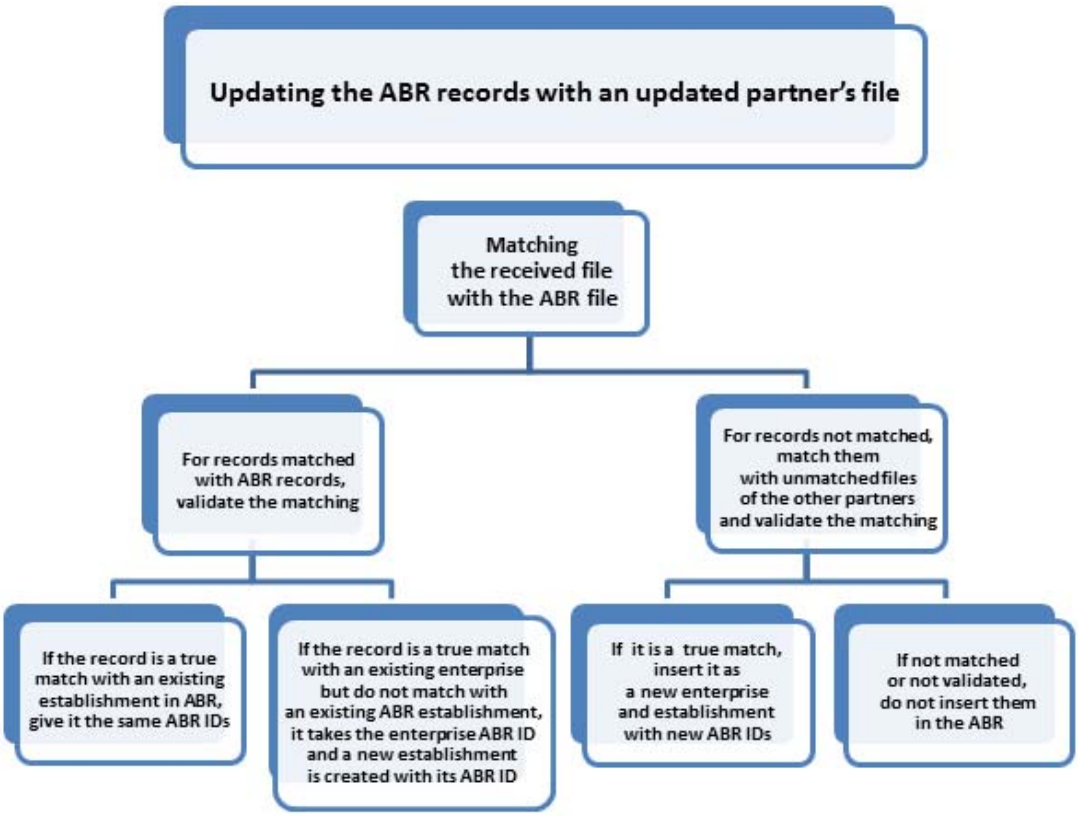
When we receive updated files from partners, we prepare them to make sure all columns are identical to what we have introduced in the database with the former version.

The first step then is to match the new records with the ABR establishment table in order to know if they correspond to an already existing enterprise or establishment. If yes, we make the proper linkages (same enterprise and establishment, same enterprise and different establishment). The ones that don't match with ABR records are then matched with the other unmatched records from all partners. The validation phase is conducted "classic" in order to insert the ones that correspond to new establishments and new enterprises.

So, the process of updating partner files to be implemented inside ABR includes three basic cases:

- A partner' record corresponds to an enterprise and an establishment not yet created in the ABR tables
- A partner' record corresponds to a new establishment of an already existing enterprise in the ABR tables
- A partner' record corresponds to an enterprise and an establishment already created in the ABR tables, just bringing a linkage for this partner that was not yet done.

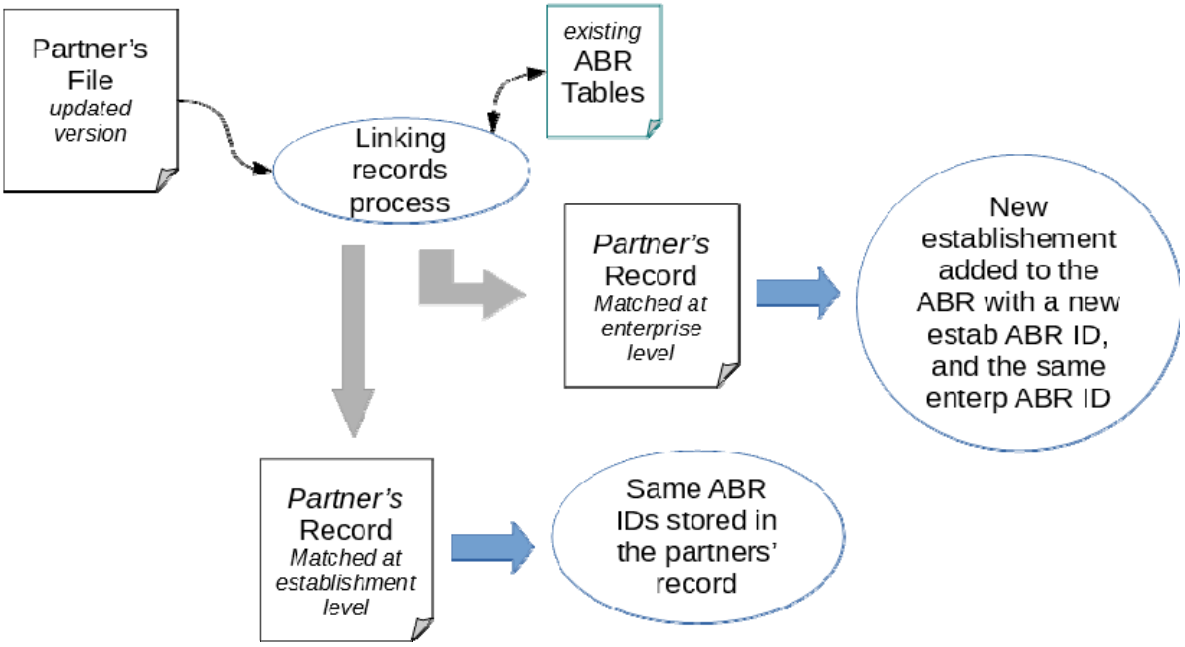
The following Diagrams Summarise the Process



Update Process

Insertion of new ABR units, update of best values for variables

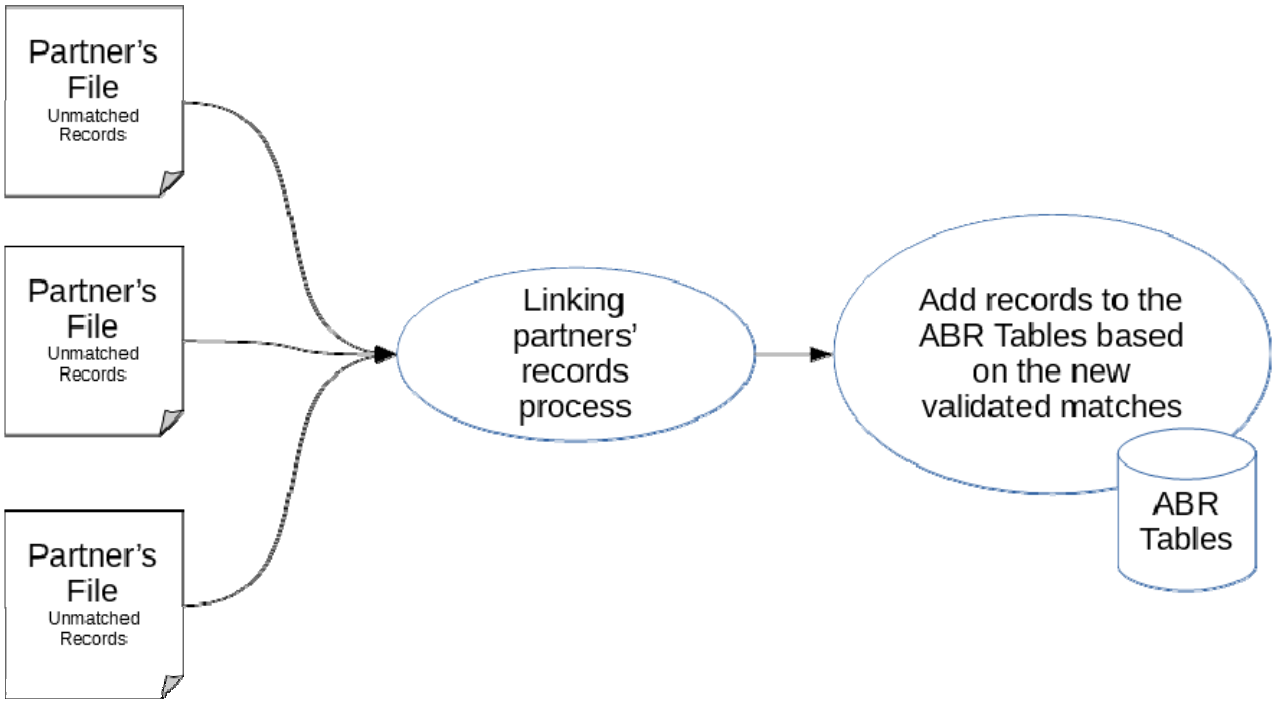
Matched Records



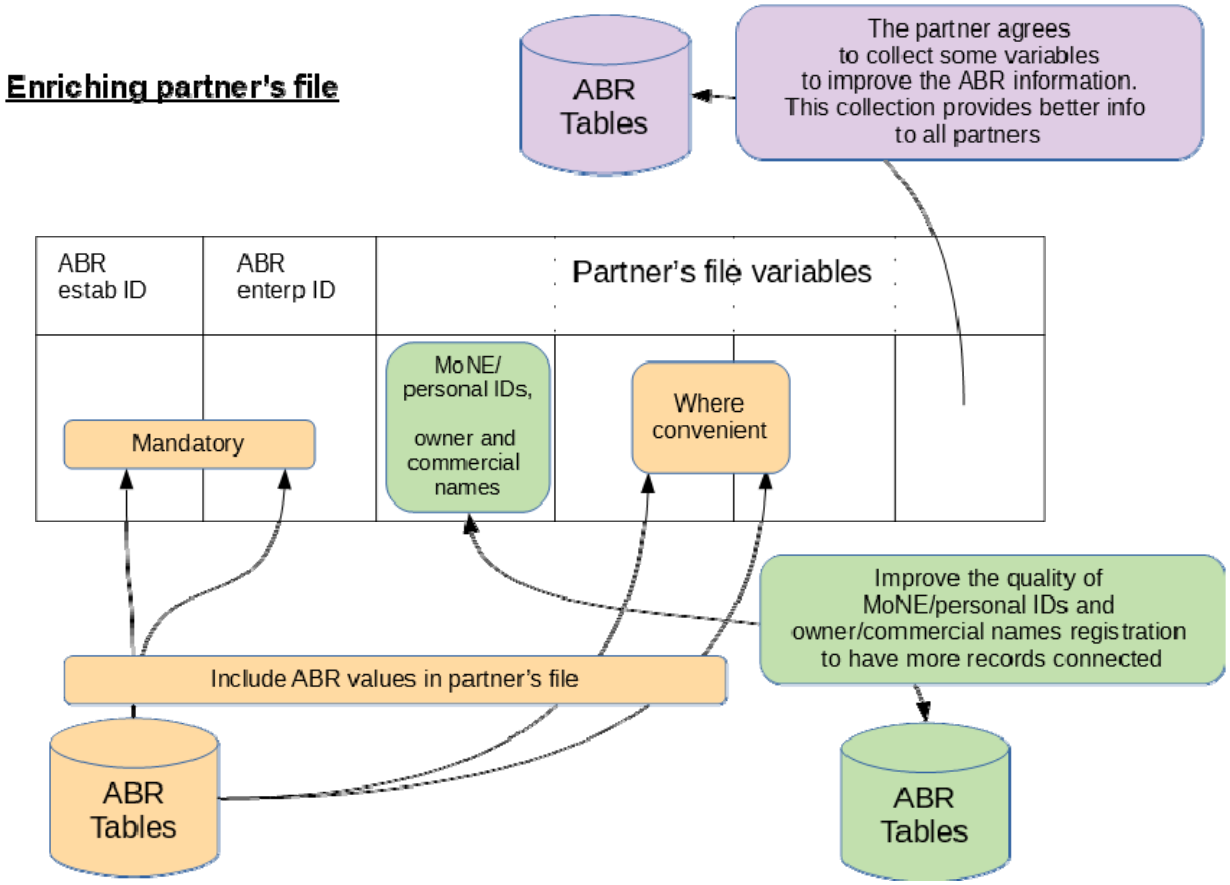
Update Process

Insertion of new ABR units, update of best values for variables

Unmatched Records



Enriching partner's file



Appendix 5: "Description of Partner's files"

Partner	Comments
Ministry of National economy (MoNE)	▪ There is no duplication in the register. ID is unique for each company (legal unit). ▪ There are no branches (the register based on the enterprises not establishments). ▪ The address registered has not to be the effective place of the activity and it happens often to be different from any effective location
Chambers of Commerce (CC)	▪ There is no duplication in the national register. ▪ In certain governorates is registered not the enterprise but an important local branch.
Municipalities (MUNIC)	▪ Municipalities' files register activity for the establishments (local units) active on their territory. So there can be duplication of establishments in several records for the several activities pursued. ▪ The addresses when registered are reliable. ▪ All the branches located on the Municipality territory are inserted in its file (field work is conducted systematically). ▪ Linkage between establishments of a same enterprise, located in different municipalities is not provided. ABR tables are a value added for that. ▪ Personal IDs in the municipalities files, when registered, are reliable
VAT	▪ The registration unit is the taxpayer. There is one record for each VAT collected. ▪ It registers some units (taxpayers) with physical address. ▪ It registers some units (taxpayers) with no physical address. These units are not registered in the ABR. ▪ It is optional to the units (taxpayers) to register several activity locations separately or under one record. ▪ To fully benefit from VAT information, it is needed to get from VAT the personal ID or MoNE ID for each record. The Tax ID (combination of personal, MoNE and Internal-to-Vat IDs) is not sufficient.

Short description of partners' files characteristics of interest for ABR management

1. Ministry of National Economy (MoNE)

- This file is exhaustive for all companies (including not active companies as it is not updated in case of end of activity)
- There is no duplication in the ID.
- There are no branches (the register based on the enterprises not establishments).
- In some cases, there is a duplication in the commercial name with different IDs, the reason is that when a company changes its legal status will be given a new MoNE ID without changing its commercial name. So both registrations correspond to a same "physical" company with two successive statuses. When choosing the active one, we look to the registration date and keep the newest one.
- The address and activity registered in the MoNE file are not reliable to be considered the effective address of the activity or the effective activity as MoNE is not following the effectiveness of the values for these two variables.

2. Chambers of Commerce (CC)

- Not all establishments of an enterprise are registered in the national file of the chambers of commerce files. So CC is not a reliable source for establishments' registering and updating.
- The file contains not only the active records; the same establishment can be registered more than once according to its status.
- The address variable is not reliable. There are many "mistakes" in registering the addresses.

3. Municipalities (MUNIC)

- The registration in the municipalities files is recording establishments (local units) with their various activities. Therefore, establishments can be duplicated when each of its activities is recorded separately.
- The addresses are reliable.
- All establishments are inserted in the files (when the municipality is going to the field to complement what is recorded at the devoted office.
- A weak point to be addressed concerns the quality of recording of Personal and MoNE IDs. Too many are missing in the MUNIC files

4. MoFP-VAT

- Taxpayers are recorded.
- When a taxpayer has to pay several VAT taxes, we find duplicated records of him.
- There are a lot of spelling mistakes.
- In many cases, the address is sufficiently clearly recorded for the info to be reliably used.
- Taxpayers records can correspond to enterprises or to establishments as the taxpayer can pay his duties in an aggregated or disaggregated way.

Appendix 6: Matching and Validation Process

Matching process

For companies, the MoNE file is the backbone file (the partners' files have their records compared to the MoNE file records).

The MoNE file was matched with VAT, Municipalities, Census 2017 and CC files using Duke tool. A first match has been done using the MoNE ID. For records not matched, a second match was done using the commercial and the owner names.

MoNE file registers enterprises. VAT file registers tax payers; each record can correspond to an enterprise or an establishment. CC file registers enterprises (and sometimes establishments). Municipalities' files record establishments (or establishments' activities). Census records establishments.

So Duke matches "automatically" records representing various units, the enterprise being the most extended one. The ABR process of validation creates ABR establishment's records and their linkage with its enterprise level.

Validate output as true matched records

The results of the automatic matching realised with DUKE has then to be validated by the validation team in order to decide the true matches, the false matches, potential duplications, and to decide the recreation of establishment records (branches, local units).

Transform a "match" proposed by Duke in a true match and decide its nature

Steps are to be followed in order to validate proposed match and give it its attributes.

1. The following has to be answered when validating records proposed match:

- Does the record correspond to an active entity?
- Does the record correspond to the level "enterprise" or to the level "establishment"?
- Is the record corresponding to an "enterprise" or "establishment" level already validated?
- How to establish the links between the establishments and the enterprise it belongs to?

Generating **ABR IDs** to create enterprise and establishment records identifiers and their links inside the ABR and with partners records is the issue to be efficiently tackled.

2. All records in the partners' files have a key identifier (a record number) that indicates its **origin position** in the **source** file as shown below:

Mone_2894	562155325	0#0#00	علا وشركاه لتاجير السيارات
|_VAT_178963	562155325#		الحزماوى لتاجير السيارات
|_VAT_178962	562155325#		الحزماوى لتاجير السيارات

The Duke output when matching the two files, here MoNE with VAT, contains a reference or a pointer in the first column that indicates the source and its record number.

3. The algorithm presents the needed variables from each file (source) for each record which allows to decide the real relation between the proposed match.
4. All available variables' values are stored in a template, built in a form which fits the database structure, and contains most variables values got from each source.
5. The template is stored in the database using a specific script.

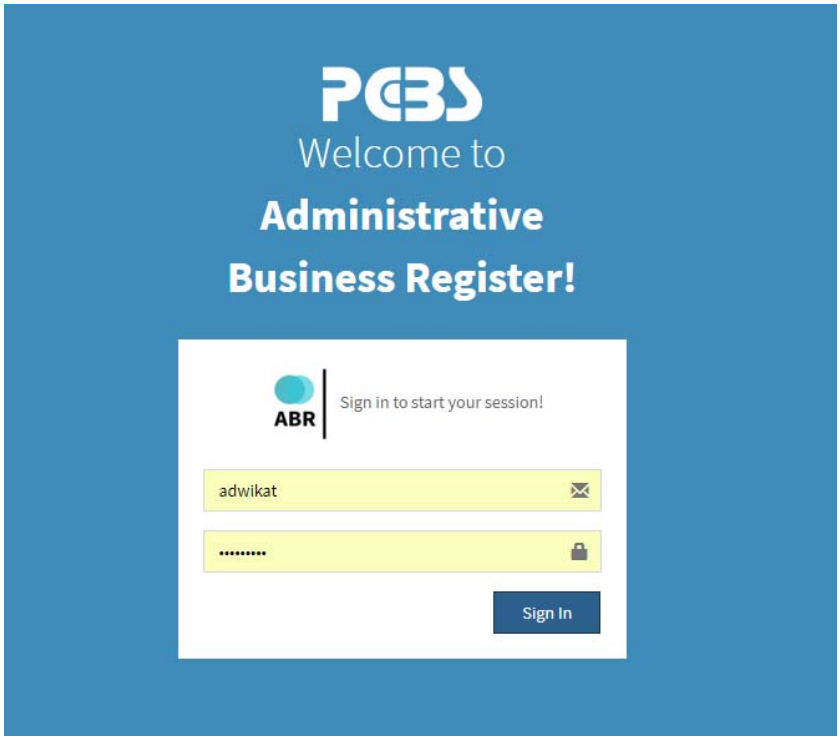
Blocks of linked partners' records in a Duke output is reproduced hereunder

	A	B	C	D	E
1	LineNumber	ID_NUMBER	COMMERCIALOWNER_NAME	TELEPHONI	ECONOMI
2	MONE_WB_7952	562116491#905	شركة خالد الحرباوي و اولاده للتجارة والأستثمار#	##	
3	|_Munic_Hebron_8338	902383215	خالد حسن رمضان الحرباوي#	0599657103 059	
4	|_Munic_Hebron_8334	902383215	خالد حسن رمضان الحرباوي#	0599657103 059	
5	|_Munic_Hebron_8333	902383215	خالد حسن رمضان الحرباوي#	0599657103 059	
6	|_chamber_4933	#902383215#90	محلات خالد الحرباوي#خالد حسن رمضان حرباوي	0599657103#22	
7	|_VAT_ALL_3	902383215	خالد حسن رمضان الحرباوي	#	
8	|_CENSUS17_WB_99974	902383215	خالد حسن رمضان الحرباوي#شركة خالد الحرباوي	022224050#059	ه مفردة
9	MONE_WB_5305	562121699#939	شركة محمد امين وأولاده للموبيليا والتجارة العامة#	##	
10	|_Munic_Ramallah_4099	562121699	شركة محمد امين واولاده للموبيليا والتجارة العامة#	#	
11	|_Munic_Ramallah_4098	562121699	شركة محمد امين واولاده للموبيليا والتجارة العامة#	#	
12	|_Munic_Ramallah_4097	562121699	شركة محمد امين واولاده للموبيليا والتجارة العامة#	2953437#	
13	|_chamber_12162	#562121699#93	شركة محمد امين واولاده للموبيليا والتجارة العامة#	#2953437#2986	
14	|_VAT_ALL_27780	562121699	شركة محمد امين وأولاده للموبيليا والتجارة العامة	#0562200200	
15	|_CENSUS17_WB_63379	562121699	شركة محمد امين واولاده للموبيليا والتجارة العامة#	022953437#056	ه مفردة
16	MONE_WB_5378	562121475#959	شركة حكم وإخوانه للتجاره العامة#حكمت عرسان	2798173##5993	
17	|_Munic_Ramallah_2566	959216144	حكم عرسان سالم سليمان#مطاعم يمي	2980343#05694	
18	|_chamber_21974	##959216144	مطاعم يمي#حكم عرسان سالم سليمان	0569444987#29	
19	|_chamber_16600	#959216151#	شركة حكمت وشركاه للأجهزه الكهربائية#	0599276326#27	
20	|_MONE_WB_5946	562172866#959	شركة حكمت وشركاه للأجهزه الكهربائية#حكمت	##	
21	|_VAT_ALL_37584	562172866	شركة حكمت وشركاه للأجهزه الكهربائية	022794869#059	
22	|_CENSUS17_WB_73612	562172866	شركة حكمت وشركائه للاجهزه الكهربائيه#حكمت	022794869#059	ه مفردة
23	|_chamber_8868	562121475##95	شركة حكم واخوانه للتجارة العامة#حكم عرسان سا	#2980343/2798	
24	|_VAT_ALL_27779	562121475	شركة حكم وإخوانه للتجاره العامة	#	
25	|_VAT_ALL_20461	959216144	حكم عرسان سالم سليمان	#	
26	MONE_WB_7995	562135251#949	شركه ام آند بي للاستيراد والتسويق#بركات أحمد	##	
27	|_Munic_Hebron_10038	562135251	شركه ام آند بي للاستيراد والتسويق#	0599889526	
28	|_Munic_Hebron_2564	949131569	بركات أحمد محمد طه#	0599889526	
29	|_chamber_2830	#562135251#94	شركة ام آند بي للاستيراد و التسويق#بركات احمد	0599889526#22	
30	|_VAT_ALL_32802	562135251	ش ام اند با استيراد وت	#	
31	|_CENSUS17_WB_95846	949131569	شركة m@b#شركة m@b#بركات أحمد طه	022250658#059	ر رئيسي يشمل

The first block is for the Company "شركة خالد الحرباوي و اولاده للتجارة و الاستثمار" that has the line number MONE_WB_7952. MoNE being the backbone for companies, the matched records from the other partners have a **"|_" tree symbol that is followed by the partners lines**.

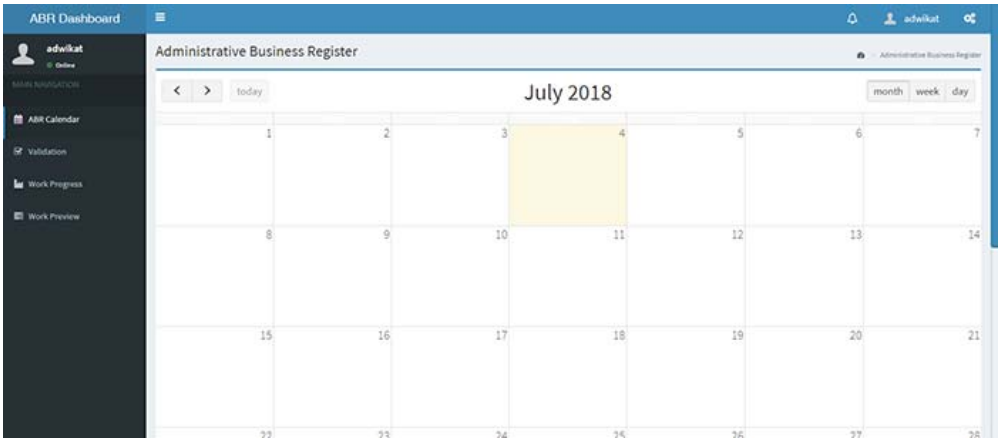
This company **MONE_WB_7952** has 6 potential records connected with it: |_Munic_Hebron_8338, |_Munic_Hebron_8334, |_Munic_Hebron_8333, |_chamber_4933, |_VAT_ALL_3 and |_CENSUS17_WB_99974.

Each line of the block has to be validated. This is done thanks to an IT interface that provides the needed information.



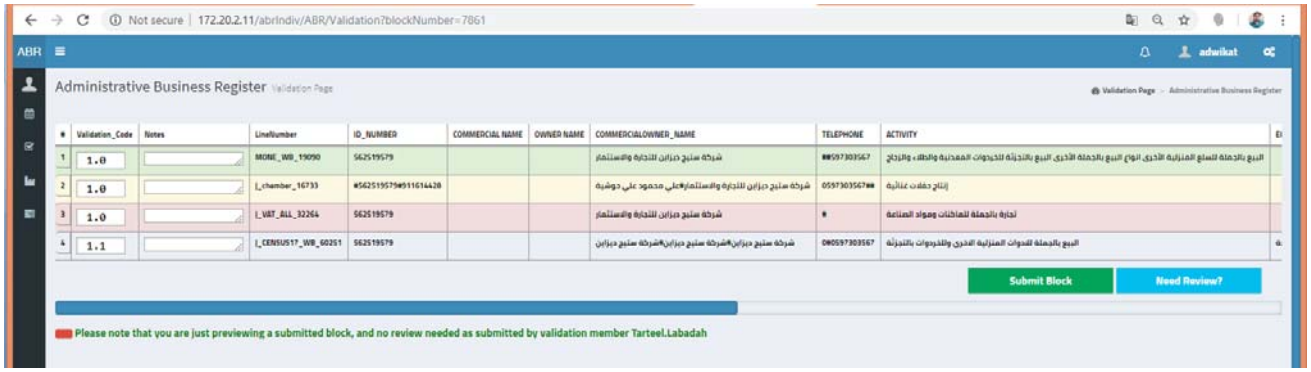
The **validation system** is a web based system designed using ASP.net. It is used by the "validation team" to validate the matching results and see which are "true matches" and at what level. Are the records matched representing the same establishment of the same enterprise? Are they from the same enterprise but different establishments? Should they be considered as false match?

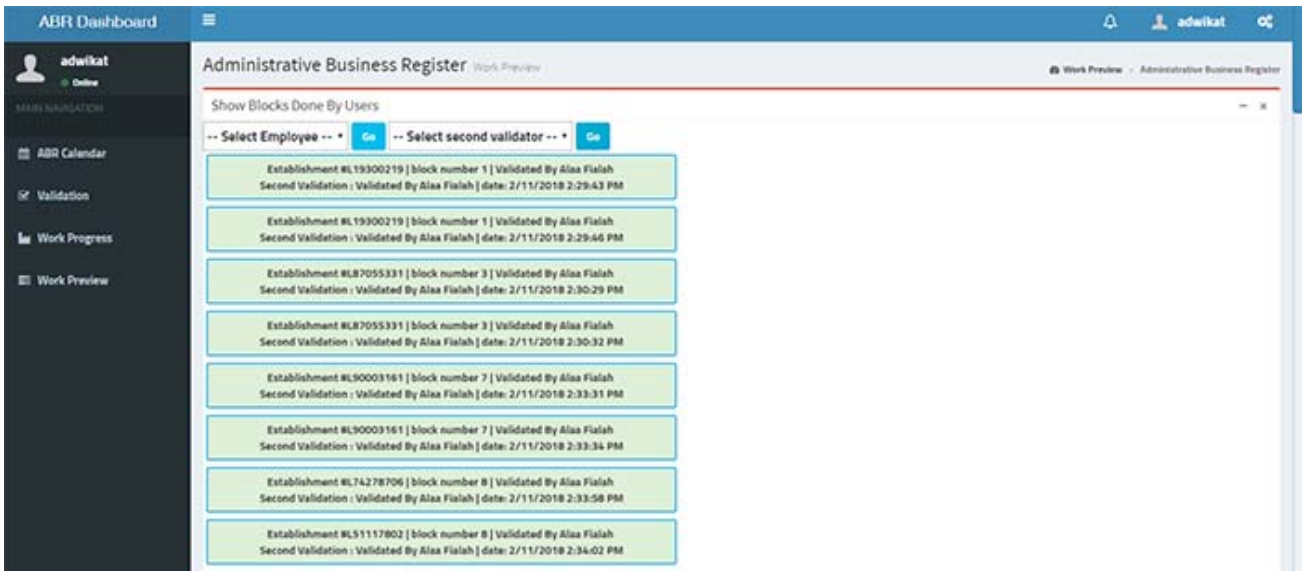
The validation system takes the data **block by block** and shows each one to be validated and submitted by a member of the validation team.



The validation process basically is manually looking at all the records in the same block and based on their variables deciding **which are true and which are false.** The validator has to inform his decision through a **Validation Code** that differentiates between the different establishments if they are more than one.

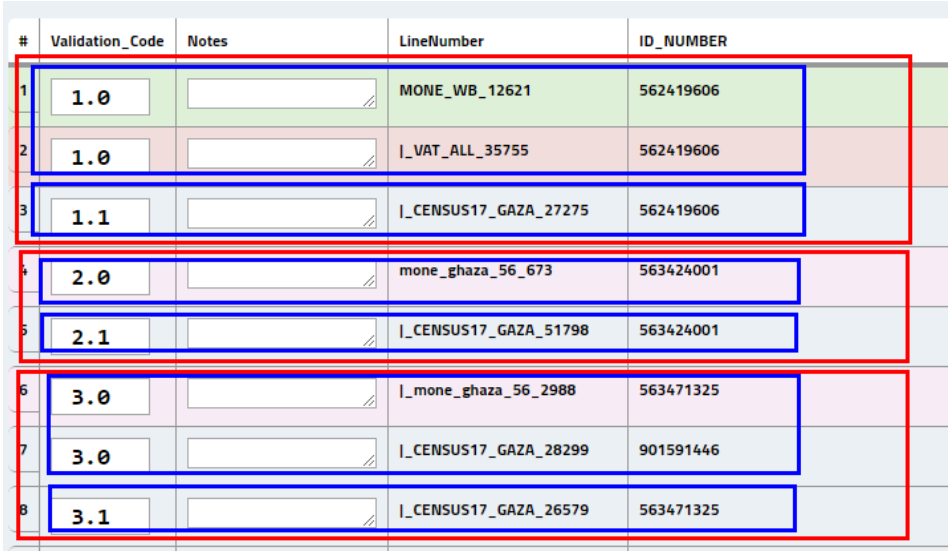
When the validator detects the first record of the block (for Companies a MoNE record) she/he should give it the code "XXX.0". The other records get the same code if they represent the same establishment. If the validator considers a record belongs to the same enterprise but not the same establishment (i.e it is another branch of the company), then she/he codes with the same integer **part of the code and different float part**, so gives it a code "XXX.1"





After completing the block validation, the validator submits the block to save the codes she/he entered in the database and that automatically generates an establishment in the establishment table and gives it a unique **ABR_ESTABLISHMNET_ID** and **ABR_ENTERPRISE_ID.**

To clarify things more let us have a look at the following table which represent 3 blocks.



#	Validation_Code	Notes	LineNumber	ID_NUMBER
1	1.0		MONE_WB_12621	562419606
2	1.0		I_VAT_ALL_35755	562419606
3	1.1		I_CENSUS17_GAZA_27275	562419606
4	2.0		mone_ghaza_56_673	563424001
5	2.1		I_CENSUS17_GAZA_51798	563424001
6	3.0		I_mone_ghaza_56_2988	563471325
7	3.0		I_CENSUS17_GAZA_28299	901591446
8	3.1		I_CENSUS17_GAZA_26579	563471325

Depending on the codes that the validator entered in the validation system, records of different partners files have been identified as belonging to enterprises and establishments (each red square represents an Enterprise, each blue square represents an Establishment)

So after submitting this block the system generates **6 establishments which are belonging to 3 different enterprises.**

The validation code number consists of two parts: the **integer** part and the **float** part. A same integer part corresponds to a same enterprise. Inside an enterprise a same float part corresponds to a same establishment.

Validation Code & ABR ID Development

The generation of ABR IDs is based on the validation code given by validators when submitted to the IT system. The validation code is converted for each line validated into an ABR establishment ID and an ABR enterprise ID.

Each line in the block has a **validation code.** The first block starts from **1.0** and is incremented by **0.1** for a new establishment of the same enterprise. **When a new block start**s, the Integer is incremented by 1, the floating value being 0 (**2.0**). A **false match** validation code is **0** for any false match line.

For example, the validation code for this simple block from different sources after multiple matching was done, should be:

Enterprise	Establishment	Status	Validation Code
Enterprise 1	Establishment 1	Branch	1.0
Enterprise 1	Establishment 2	Branch	1.1
Enterprise 1	Establishment 2	Duplicated	1.1
Enterprise 2	Establishment 1	Branch	2.0
Enterprise 2	Establishment 2	Branch	2.1
Enterprise 2	Establishment 3	False match	0

How will ABR ID will be generated based on the validation code?

First, the ABR ID have a length of **9 digits** consists of the following form:

ABR ID Enterprise Form	ABR ID Establishment Form
[E] [7-RANDOM] [1-CHECKSUM]	[L] [7-RANDOM] [1-CHECKSUM]

Where **E** indicates an **Enterprise**, and **L** an **Establishment** (Local unit). The last digit is a check sum digit generated with the previous 7 digits random numbers using **Luhn algorithm methods.** The 7 digits are generated randomly and a check verify that the number generated has never been used, neither for an ABR enterprise ID nor an ABR establishment ID, so each number will be **unique**.

For example, the following ABR ID's is associated with the records submitted, as shown below :

Enterprise	Establishment	Validation Code	ABR_ENT_ID	ABR_EST_ID
Enterprise 1	Establishment 1	1.0	E61315164	L83451237
Enterprise 1	Establishment 2	1.1	E61315164	L71784649
Enterprise 1	Establishment 2	1.1	E61315164	L71784649
Enterprise 2	Establishment 1	2.1	E40491730	L62033618
Enterprise 2	Establishment 2	2.2	E40491730	L12973384
Enterprise 2	Establishment 3	0	NULL	NULL

Each enterprise holds a unique ABR enterprise ID. Its establishments have this unique ABR enterprise ID and have an ABR establishment ID.

All enterprises and establishments ID's are connected together with a **relational table** that links the enterprise to its establishments. **One enterprise can have one or more establishments. Most enterprises have only one establishment**.

Company / Non-Company

Two types of enterprises exist that are processed specifically in the ABR: companies and non-companies.

Ministry of National economy (MoNE) is the source of registering companies. Its register is exhaustive as each company has to register if it wants to benefit from this statute. An ID is given by MoNE that consists of **9 digits** beginning with **56**. This ID is crucial for the company in all its financial transactions (import, export, tax, …). It is (should be) registered by all institutions (municipalities, chambers of commerce, VAT, …). Establishments belonging to a Company do not have a specific additional identifier.

Non-companies have no such ID. The Personal ID of the owner is the most often used identifier for non-companies. When a person owns more than one enterprise or establishment, most partners do not register specific identifiers for each.

The process of linking records has to address this lack of unique identifier for establishments.

Appendix 7: Database Description

ABR Database Structure Development

ABR database consists of **input tables** to receive partners files, where each record is stored in its own table (one for each partner), **validation table** that contains the blocks of multiple matching with all useful variables values for each record that contribute to decide the match coding, **audit table** to store all information related to any ABR modification during the update process. Those modifications consist of update, deletion or insertion of new establishments/enterprise, **Variables' source table** that indicates the origin of each shared variable from within the ABR data, and **Parties and configuration tables** to help manage the system, groups of users and users' permissions.

Implement the algorithm for inserting the matching results to database storage

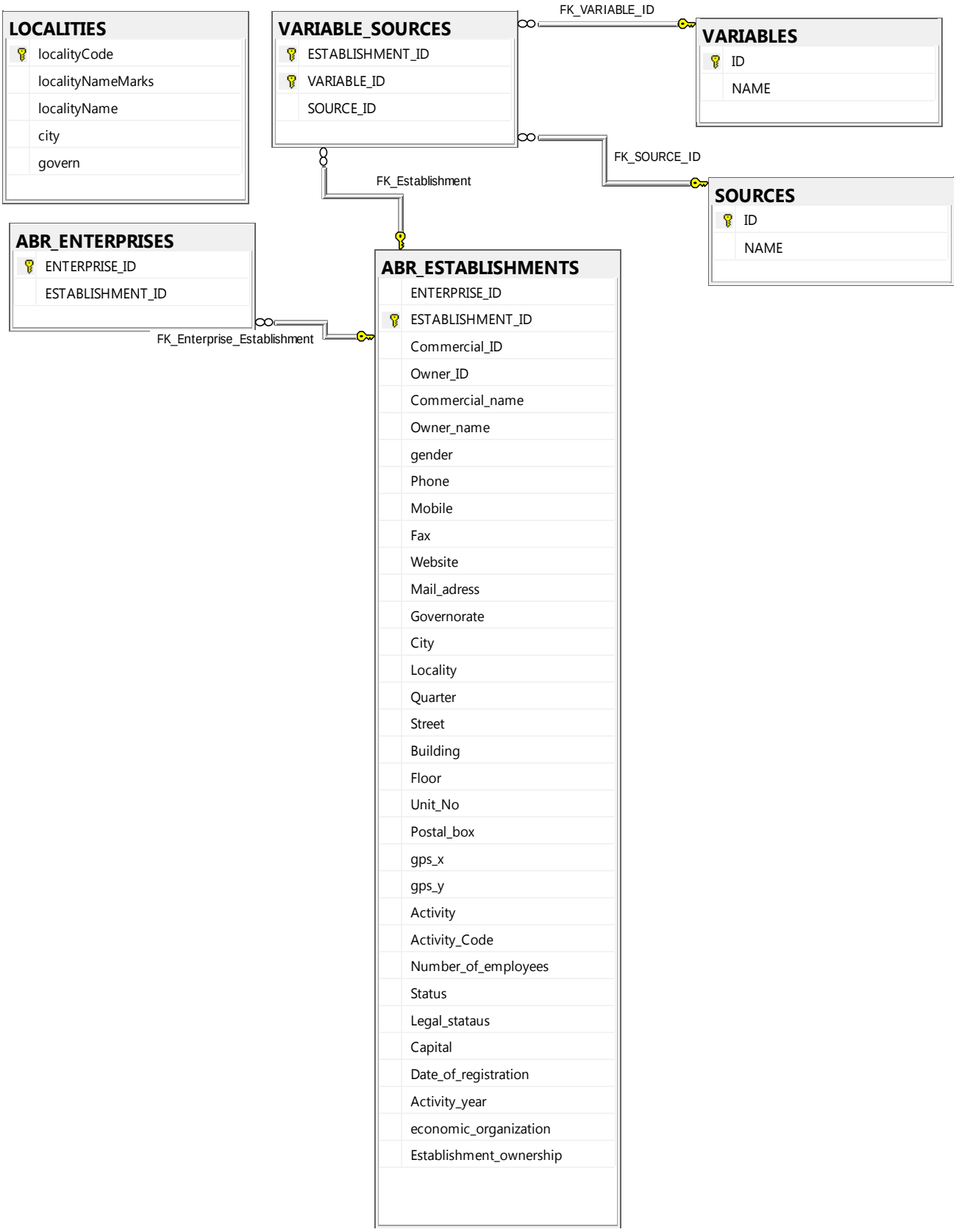
The main ABR tables' construction (ABR_Enterprises and ABR_Establishements) is mainly based on the validation data of the matching, enriched by information from the partners' tables.

This process is based on specific rules used to fetch the appropriate information from the right input.

All these rules are implemented within a T-SQL script that considers companies and non-companies' cases by respecting the sources' priorities for each case. The script is also responsible for populating the origin of each shared variables populated within the ABR tables to keep a track of the information's provenance.

Database Schema and System Structure

ABR tables schema



Partners tables

P_VAT26000
- LineNumber
- ID_NUMBER
- TELEPHONE
- COMMERCIALOWNER_NAME
- QUARTER
- REGION
- BUILDING
- FLOOR
- STREET
- FLAT
- ACTIVITY
- REGISTRATION_DATE
- CLASSIFICATION
- LEGAL_STATUS
- CREATION_DATE
- Column1
- LineNumber_num
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

census_wb
- LineNumber
- governorate_code
- governorate_name
- locality_code
- locality_name
- COMMERCIALOWNER_NAME
- OFFICIAL_COMMERCIAL_NAME
- USED_COMMERCIAL_NAME
- OWNER_NAME
- sex
- TELEPHONE
- lanline
- mobile
- website
- quarter
- STREET
- building_name
- operational_status
- activity
- establishment_ownership
- economic_organization
- legal_status
- year_of_activity
- total_No_of_employees
- preparation_of_accounting
- capital
- ID_NUMBER
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID
- GPS_X
- GPS_Y
- Strata
- locality_fromGIS

munic_Nabuls
- identifier
- COMMERCIALOWNER_NAME
- OWNER_NAME
- Activity
- address
- date_of_registration
- COMMERCIAL_NAME
- ID_NUMBER
- telephone
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

census_ghaza
- LineNumber
- governorate_code
- governorate_name
- locality_code
- locality_name
- COMMERCIALOWNER_NAME
- OFFICIAL_COMMERCIAL_NAME
- USED_COMMERCIAL_NAME
- OWNER_NAME
- sex
- TELEPHONE
- lanline
- mobile
- website
- quarter
- STREET
- building_name
- operational_status
- activity
- establishment_ownership
- economic_organization
- legal_status
- year_of_activity
- total_No_of_employees
- preparation_of_accounting
- capital
- ID_NUMBER
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID
- GPS_X
- GPS_Y
- Strata
- locality_fromGIS

cc17_wb
- LineNumber
- COMMERCIAL_NAME
- MANAGER_NAME
- commercial_owner_name
- mobile
- phone
- fax
- telephone
- ID1
- ID2
- ID3
- ID_NUMBER
- street
- city
- Activity
- file_status
- number_of_employees
- creation_date
- registration_date
- website
- email
- sex
- [Ent No.]
- legal_status
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID
- ownerID_selected
- commercialID_selected
- ownerName_selected

munic_Hebron
- LineNumber
- COMMERCIALOWNER_NAME
- COMMERCIAL_NAME
- ID_NUMBER
- OWNER_NAME
- classification_No
- classification
- Activity
- building
- STREET
- TELEPHONE
- unit_No
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

munic_Salfit
- identifier
- COMMERCIALOWNER_NAME
- Activity
- locality
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

munic_Qalqilia
- identifier
- ID_NUMBER
- COMMERCIALOWNER_NAME
- COMMERCIAL_NAME
- date_of_license
- street
- OWNER_NAME
- Activity
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

munic_Jenin
- identifier
- COMMERCIALOWNER_NAME
- ID_NUMBER
- address
- Activity
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

mone_ghaza_56
- LineNumber
- Serial_number
- ID_NUMBER
- COMMERCIALOWNER_NAME
- legal_status
- registration_date
- city
- address
- telephone
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

Mone_wb
- LineNumber
- governorate
- COMMERCIAL_ID
- ID_NUMBER
- COMMERCIALOWNER_NAME
- COMMERCIAL_NAME
- partners_names
- partners_IDs
- legal_status_code
- legal_status
- registration_date
- capital
- currency
- city
- address
- street
- building
- telephone
- phone
- fax
- mobile
- postal_box
- EMAIL
- status_code
- status_description
- activity
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

munic_Tubass
- LineNumber
- COMMERCIALOWNER_NAME
- COMMERCIAL_NAME
- OWNER_NAME
- Activity
- status
- Nature
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

munic_AlBirah
- LineNumber
- COMMERCIALOWNER_NAME
- address
- OWNER_NAME
- Activity
- COMMERCIAL_NAME
- status
- TELEPHONE
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

vat_all
- LineNumber
- ID_NUMBER
- COMMERCIALOWNER_NAME
- quarter
- region
- region_code
- STREET
- building
- floor
- flat
- Activity
- email
- box
- Fax
- website
- TELEPHONE
- landline
- Mobile
- registration_date
- classification
- legal_status
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

fichier_ghaza
- LineNumber
- identifier
- COMMERCIAL_NAME1
- COMMERCIAL_NAME2
- COMMERCIALOWNER_NAME
- city
- locality
- street
- Activity
- OWNERSHIP_DESC
- registration_date
- No_of_employees
- landline
- mobile
- TELEPHONE
- postal_box
- website
- email
- capital
- currency
- legal_status
- operational_status
- total_No_of_employees
- male_employees
- female_employees
- LICENSE_ID
- the_source_of_the_license
- source_of_the_license
- ID_NUMBER
- partners_IDs
- partners_names
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID
- ownerID_selected
- commercialID_selected
- ownerName_selected

munic_Ramallah
- LineNumber
- ID_NUMBER
- COMMERCIALOWNER_NAME
- TELEPHONE
- OWNER_NAME
- COMMERCIAL_NAME
- Activity
- phone_1
- Phone_2
- activity_code
- Activity_in_details
- BUILDING_No
- STREET_CODE
- STREET
- unit_No
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

munic_Tulkaram
- LineNumber
- identifier
- COMMERCIALOWNER_NAME
- MANAGER_NAME
- city
- postal_box
- TELEPHONE
- Activity
- account_date
- street
- ID_NUMBER
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID

munic_Bethlehem
- LineNumber
- No_of_file
- ID_NUMBER
- COMMERCIALOWNER_NAME
- OWNER_NAME
- street
- Activity
- COMMERCIAL_NAME
- phone_No
- mobile
- TELEPHONE
- date_of_license
- governorate
- ABR_ENTERPRISE_ID
- ABR_ESTABLISHMENT_ID
- ownerID_selected

Appendix 8: Data Flows

Discussing the partnership content with each Partner:

The partners collect data according to their needs, following the laws that apply to their tasks. The ABR System has to make the better use of their data.

The ABR System needs to be provided by the variables which are collected by the partners that can inform directly the shared variables of the ABR and also other variables that help deciding the true match at enterprise and/or establishment levels.

The quality of the ABR reflects the quality of the data registered by each partners. Some variables need highest attention because of their role for the matching process or because they will be used to decide the best values for shared variables.

Also the format in which the partner provides (later, uploads) its data is a key question for securing an efficient work process.

Example: the MoFP-VAT data.

For the V1, the format in which the file has been made available implied important manual treatments in order to construct a usable Excel file. Such a prototype process should not be repeated for several reasons, one being the risk of deteriorating the information.

The MOFP team responsible for the data has proposed to manage the issue of the format in which data will be transmitted with the general directorate of Tax Computer.

Also discussions include variables to be made available. For each record, the ABR System can secure proper matching when both ***Personal ID and VAT ID are made available (if fact, the table of connections between both IDs should be provided for the V2 onwards)****. Also, exchanging technical knowledge should help manage duplicated records, dates of creation and modification of some variables.*

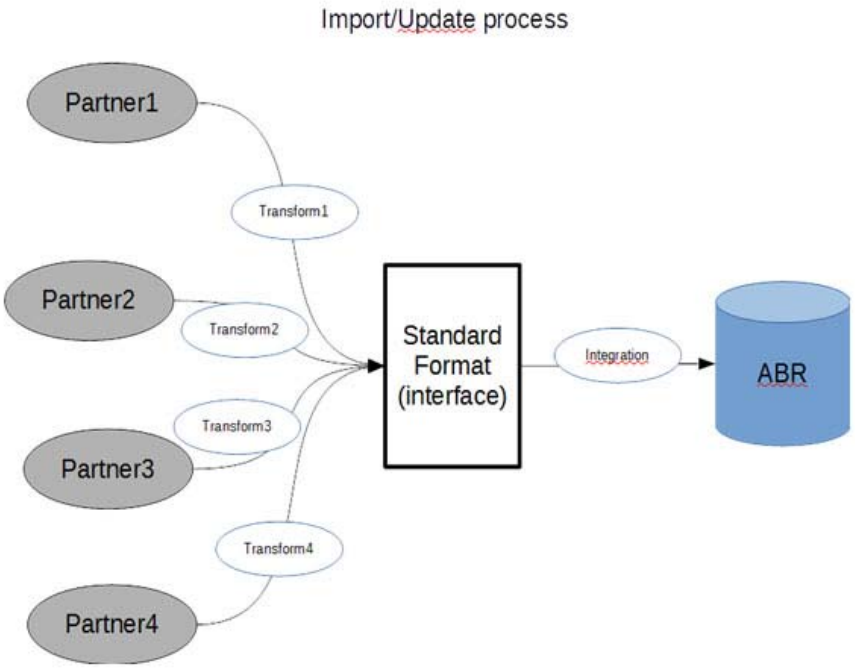
Each progress will have dramatic positive effect on the quality of the ABR System.

Expected automatic exchange of data between partners and the ABR System

When the IT tools will be available, the partners will upload and download data using the ABR system.
NB: The System will secure the integrity/confidentiality of each partner data as these data are not public except the shared ABR variables (see below).

Each partner will benefit from a script in order to integrate its data in the ABR System. Adding a script to the partners system will reflect a “standardized” template of import and export from and to the ABR adapted to both Partner’s and ABR’s IT systems.

Each partner should import to the ABR system based on an agreed Standard format and each partner can export from ABR and import to his system what he needs to enhance his information system according to the quality of ABR and the best variables included in the ABR System (export scripts needed for the partners).



The imported data processed in the ABR System will respect the **confidentiality** of the information treated**.** Only values that correspond to the ABR shared variables (see the list above) will be disseminated to the partners of the System through the ABR tables.

At the stage of building the ABR tables (V1 and future V2), the values for shared ABR variables existing in partners' files are treated automatically in order to choose the best available values taking in consideration the standard quality of the data collection.

In the permanent functioning of the ABR System, each partner could be specifically asked to collect some variables or check values of some variables in order to provide new information and improve dubious/low quality values.

Using a readable format (interface) for providing data

For each file provided (initial or updated), it is needed that the partners provide the agreed set of variables in the agreed format (Excel file, for instance).

The process of updating should benefit of continuities as much as possible. For instance, the provided file should keep the same names for each variable than in former versions provided.

Work put on partners should be minimized, so the ABR System should accept the format provided for its file. As experience has shown, there can be a limit to the format acceptable (ex. VAT file structure), which then has to be discussed sorely in order to be provided with a standard-readable form.

The quality of the recording of variables varies from a partner to another and also between records of a same file. The participation to the ABR System constitutes an invitation to collect data better. Also some partners are well positioned to collect data they don't register to date (ex. Collecting GIS coordinates for municipalities).

New partners will be integrated in the System. Based on the experience gained with constructing the ABR-V1, smooth discussions can be followed with the new partners in order to help them prepare the files they provide and develop the script to integrate the partners system with ABR system.

The export phase should be added in a scheme.

Appendix 9: Future System

The general features for the improvement aimed with V2 and V3 are described in Part 4 of this document.

This appendix develops proposals for some IT aspects intended to be in use in 2020+.

Graphical user interface (GUI) for partners:

The GUI should be user friendly and meet the best user experience and provide easy and efficient way for partners to the System to process all functions

As a reminder, basics requirements are browsed below, that will be discussed in full details:

1- Login page
2- Register page (with approval)
3- Main page (dashboard): the dash board will NOT be the same for each user , each user should have the dashboard that meets his various needs, if he is a data provider or admin of the ABR System or Supervisor for matching validation, and so on.
4- Browsing establishment/enterprise tables: include full variables filters and search with all the possible permissions (advance search and filter)
    - The page should include all the basic functions (trash, update, insert) depending on who is the user and to which permitted access he is entitled
    - Links for all features (reports , printing, statistics, matching, etc.)
5- Matching page: browsing the updated records (this point should be discussed in more details since it is a crucial page, core to the ABR quality). It will include matching operation with the programming of dynamic distance measurement in order to fine tune the proposed linkages, validating process to decide true and false matchings.
6- Establishments page
7- Enterprises page
8- Partners pages (one page for each partner)
9- History and logs pages
10- Full dynamic filtered counting to get statistics for each user needs
11- Reports page: presenting outputs along criteria agreed with partners for shared variables. Printing and publishing functions should be available (excel , csv, images , graphs)
12- Users and permissions pages: to create , edit , delete all type of given permissions
13- Search and advance search page: searching using various criteria including dates, names, IDs , ABR IDs.
14- Information pages: project history, project details, partners, general info, contact, mission and vision, user guide, assistance to user, privacy and use agreements, etc.

Frequency of providing data from partners to update the ABR tables

The frequency of receiving data from partners shall be adapted to each partner in connection with his own process of updating his register.

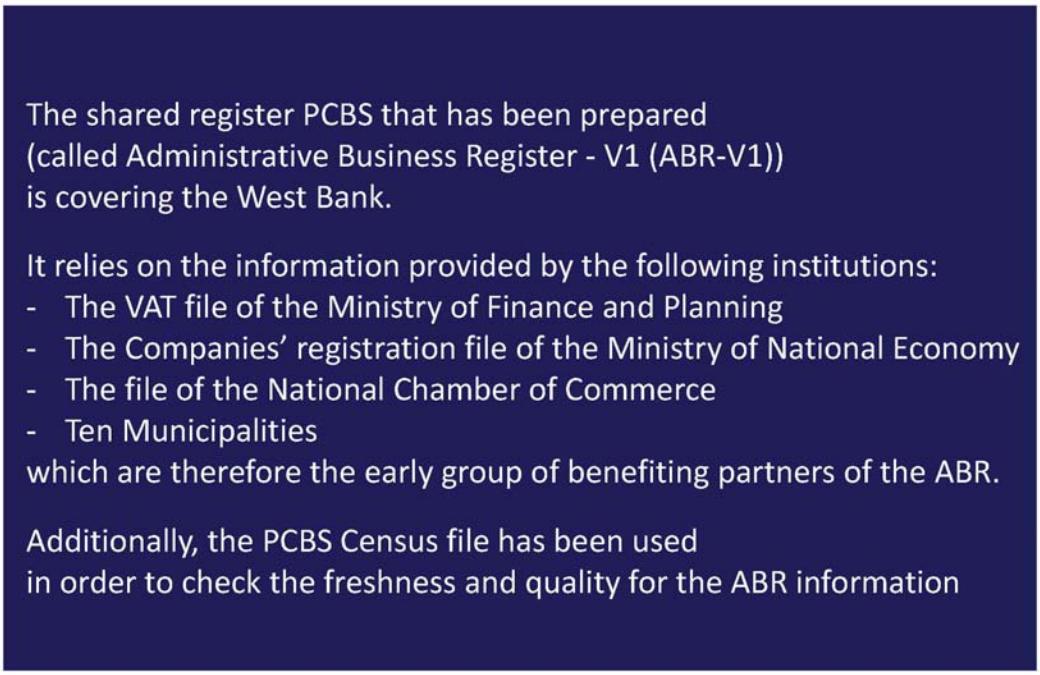
Creation of new records and updating information of existing records will follow different frequencies also in consideration to the processes for ABR updates.

As a standard rule (to be adapted where appropriate), the uploading of partners' new records could be done monthly, quarterly or even yearly.

The updating on existing records values with partners' information could be considered at different frequencies. Preliminary hypothesis are: monthly or quarterly with VAT, yearly with Municipalities and CC.

MoNE information for new created companies could be provided monthly. As MoNE doesn't update the values it registers for shared variables, it would not be asked to provide its file for updates. But it would receive the updated ABR tables in order to have access to best values' available.

Appendix 10: PowerPoint synthetic presentation of the achievements and perspectives



BUSINESS REGISTERS' DEVELOPMENT IN PALESTINE
SITUATION FOR THE WEST BANK ABR-V1 – DECEMBER 2018

- Each partner administration has its own register (some have even several ones).
- The ABR aim is to provide each partner with data that help contacting the economic units it registers (the identification data), in order to update its own register:
  - to add units it should have but did not register yet,
  - to benefit from up-to-date values for the variables that identify each economic unit (in particular MoNE ID, Personal ID, commercial/owner names, address),
  - to be able to link in its file the local units (i.e. establishments, branches) that belong to the same single enterprise, company.

- The linkages between the ABR economic units and each partner's units are made possible through IDs that are generated by the ABR System. These IDs are "internal" to the system of exchange of data between the ABR partners (they are not for public use):
  - an Enterprise ABR ID that connects each enterprise, company, registered in the ABR file with its counterpart(s) in the partner file
  - an Establishment ABR ID that connects each local unit (i.e. establishment, branch) registered in the ABR file with its counterpart(s) in the partner file.
- Both IDs are linked together in order to provide a global overview of the enterprise and its establishments.

BUSINESS REGISTERS' DEVELOPMENT IN PALESTINE
SITUATION FOR THE WEST BANK ABR-V1 – DECEMBER 2018

The ABR follows some basic rules that makes it unique for partners:

- It contains variables useful to identify active economic units
- It registers enterprises, companies, and local units (establishments, branches). It connects each enterprise to its establishments
- It has no duplication between the units it registers. Each enterprise, each establishment, is registered only once (with one record).

This point is crucial. A unit is introduced in ABR only when there is an evidence it is active AND it is not already registered in the ABR. Without sufficient evidence, the introduction of a unit is postponed.

As a consequence, some active units recorded in partners' files are not yet included in the ABR. Improving values of variables collected by partners should provide needed evidences to add records in the ABR.

The West Bank records in the ABR

The 2017 Census has registered for West Bank (through field work) almost 100 000 economic establishments.

If large units are usually registered by one or more partners, small ones are often not registered or badly registered.

The ABR connects information from partners' files corresponding to 37 000 active establishments

- The almost 31 000 Enterprises in the ABR table for WB corresponds to these 37 000 Establishments
  - 14 000 companies corresponding to 18 000 establishments
  - Almost 17 000 non companies corresponding to 19 000 establishments.

BUSINESS REGISTERS' DEVELOPMENT IN PALESTINE
SITUATION FOR THE WEST BANK ABR-V1 – DECEMBER 2018

ABR companies' records for WB linked with partners files

Enterprises table	ABR Companies								
Mone_WB									*14 041*
VAT									*11 815*
Chamber									*4 206*
Municipalities									*3 682*
Total	*1 356*	*213*	*1 616*	*1 957*	*6 911*	*680*	*522*	*786*	**14 041**

- The above table concerns companies registered in the West Bank
- 14 000 WB companies are included in the ABR enterprise table
- Beyond being in the MoNE file (that registers all companies), a company has to be at least registered in a second partner file.
- This second registration is used as an evidence the company is active.

The table reads as follows:
- 1356 companies are registered in all the partners files;
- 1591 are registered with all partners files except CC file;
- 786 companies are not registered in partners files but the info about its activity has been checked with the Census.

ABR records for non companies in WB

Enterprises table	ABR Non-Companies (WB)							
VAT								*9 891*
Chamber								*2 363*
Municipalities								*11 355*
Total	*1 002*	*4*	*3 429*	*1 319*	*4 131*	*38*	*6 920*	**16 843**

- Linkages with VAT, municipalities and Chamber of commerce files have allowed to include almost 17 000 active enterprises.
- The linkages for non companies is more difficult than for companies because of lack of proper values for "personal ID" and "owner name" variables in many partners' files (there is a lack of precise addresses for where activity is carried out).
- A major issue for the future work will be to progressively improve the quality of these information in partners' files.

ABR companies' records for Gaza

Matching records of the MoNE Gaza file that are not included in the MoNE WB file allowed to add some 2200 active companies registered there (checked with other partners' files).

ABR non companies' records for Gaza

- The information provided for Gaza by Ministry of Labour is a listing of 55 000 economic units, a size very much in line with the 65 000 establishments registered in Gaza during the 2017 Census.
- The recorded values for the variables used to introduce a unit in the ABR do not provide enough evidences for decision.
  Nevertheless, this promising file should allow further development (more extensive recording of MoNE and Personal IDs, owner names…) in order to be able to securely integrate units in the ABR tables (especially non-companies).

Sharing the ABR Version 1 information

Each partner shall receive:

The ABR tables including all the economic units already registered (with all identification variables that allow contacting the unit).

- The file that each partner has provided. Each linked record will include now the two ABR IDs generated. They will allow a smooth management of further exchanges with the partners for each update of the ABR records and linkages with partners'.

In order to fully benefit from the ABR, now and in the future,

- Each partner is expected to update its files:
  - To introduce in its IT system the two ABR IDs for its connected records in order to manage exchanges between its file and the ABR
  - To consider its variables' values that are different from the ABR ones. To signal doubts about ABR values
  - To add ABR records that are lacking in its file (along its aims scope)
- Each partner is expected to register some information with the highest quality in its new records:
  - The update will be of utmost quality if each partner registers systematically with care four variables: MoNE ID, Company name, Personal ID, Owner name
  - Also it is desirable to improve the address registration (GIS when available, as detailed as possible other while)

Four situations are of particular interest
to look at ABR value-added and further progress to achieve

- Records for which the variables' values in the ABR and in the partner file are identical (the partner information has been taken to provide the values in the ABR lines). Then ABR provides no value-added.
- Records for which the variables' values in the ABR are different from the values in the partner file. Then ABR provides a value added and allows the partner to update its information.
- Records for enterprises that are in the ABR but not in the partners' file. The coverage of the partner file can be improved.
- Records for which the ABR registers more establishments than the partner. The coverage of the partner can be improved.

Examples will be provided to each partner
in order to show such cases and to discuss what to do next

Appendix 11: PowerPoint of the presentation of the ABR at the final event of the project

Palestinian Business Register

Insee EXPERTISE FRANCE

Palestinian Business Register

Insee EXPERTISE FRANCE

The Objectives of the Business Register

A register that is updated regularly for all operating economic units which contains identification data for all partners

Ensuring regular updating

Providing links among the various administrative business registers data to produce a centralized business register that provides tools for improving the coverage and quality of data of economic units

Increasing the inclusiveness and improving the quality

The Partners of BR–V0

10 Large Municipalities (Munic)

Union of Chambers of Commerce (CC)

Ministry of National Economy (MoNE)

Ministry of Finance and Planning

-

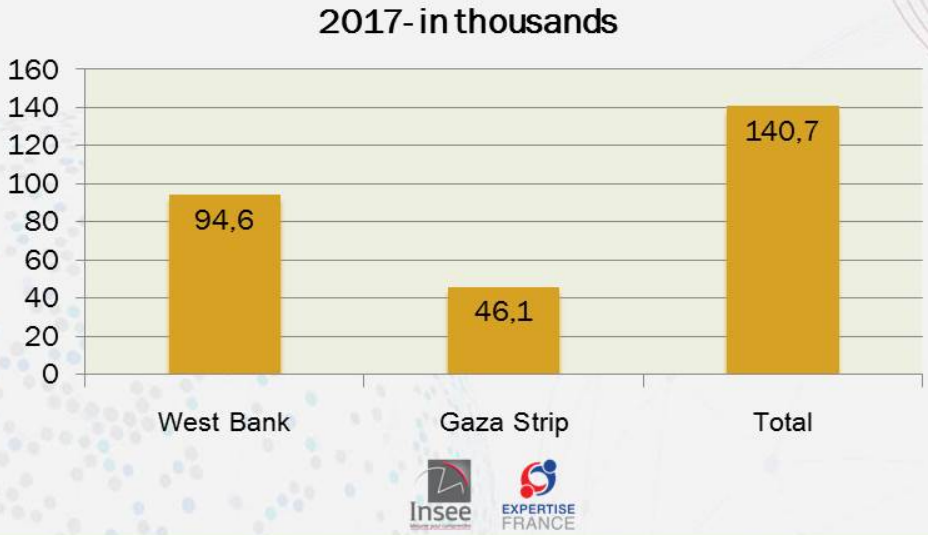
Value Added Tax (VAT)

The Role of Business Register

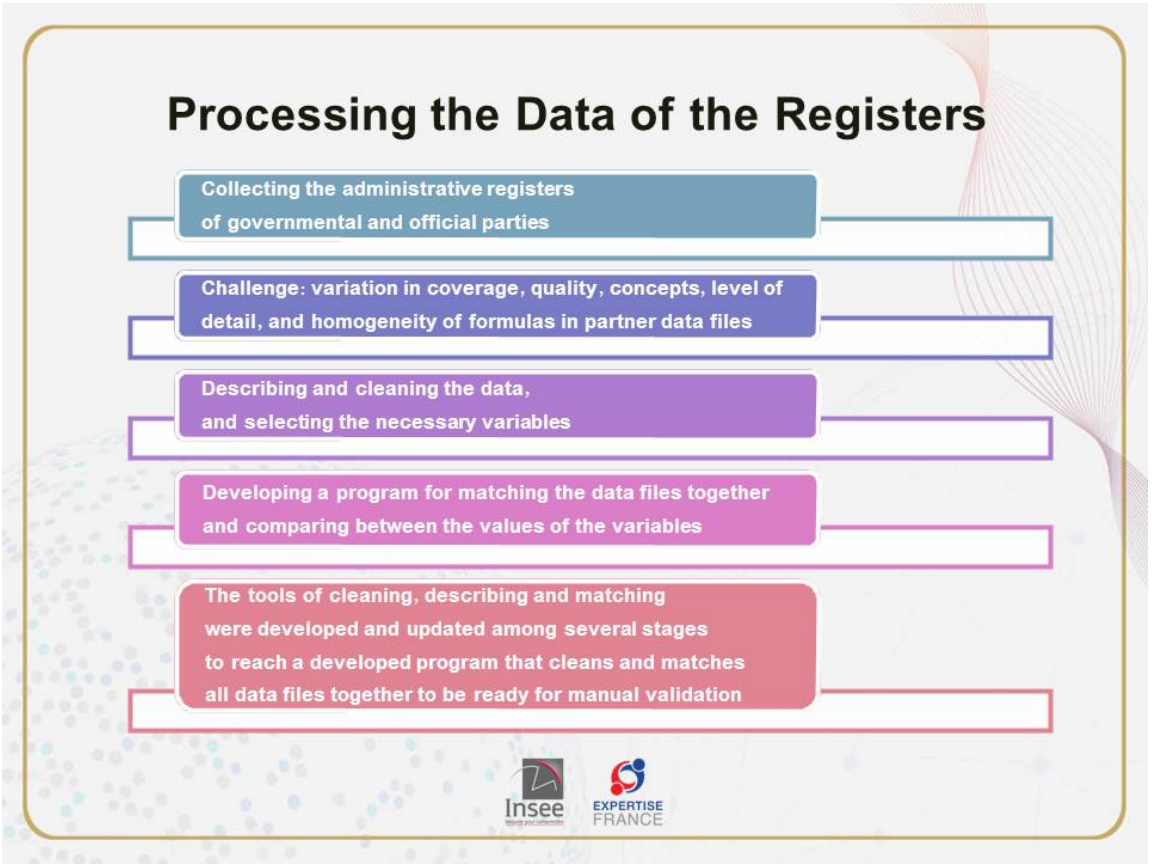
Insee EXPERTISE FRANCE

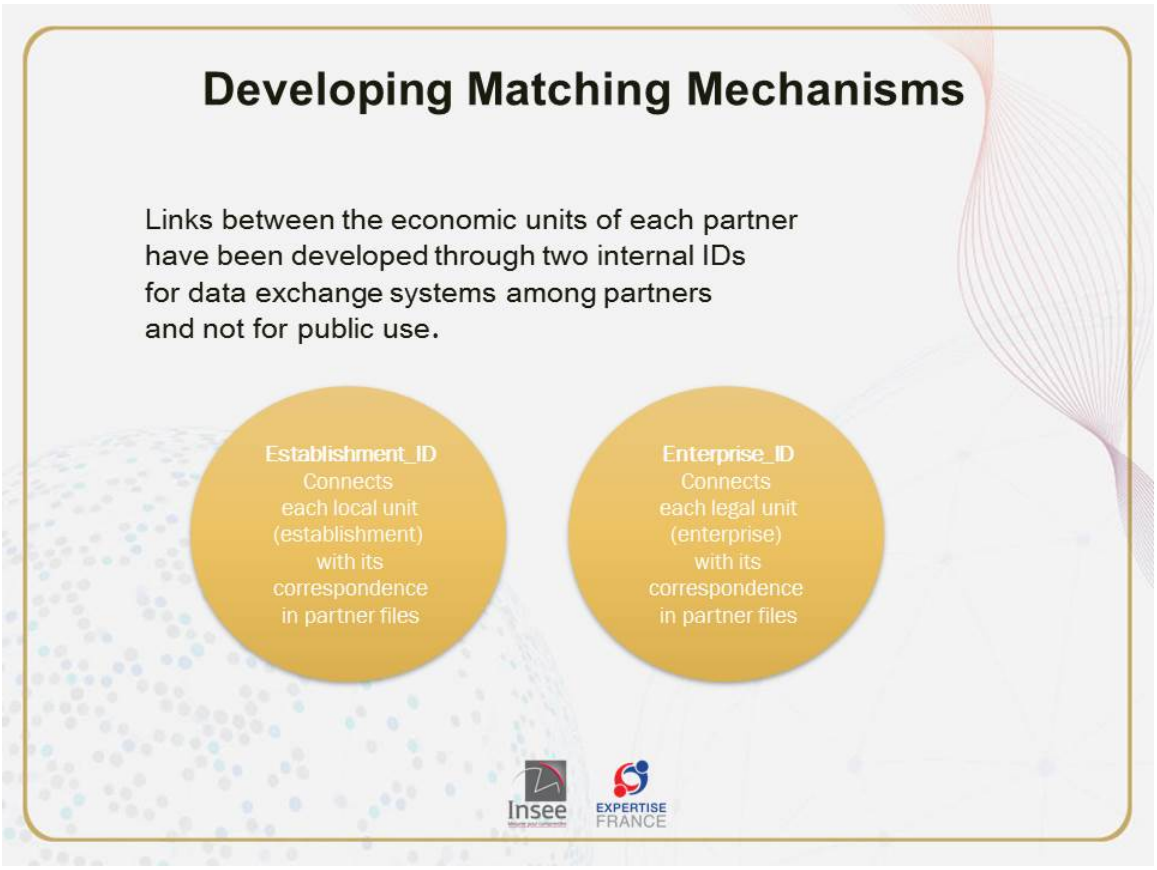
The Current Situation in Palestine

The main source of statistical information about establishments and enterprises is the establishment census which is carried out every 5 years by PCBS.
The latest census was used to examine the comprehensiveness and quality of administrative records



Insee EXPERTISE FRANCE


Processing the Data of the Registers
Collecting the administrative registers of governmental and official parties
Challenge: variation in coverage, quality, concepts, level of detail, and homogeneity of formulas in partner data files
Describing and cleaning the data, and selecting the necessary variables
Developing a program for matching the data files together and comparing between the values of the variables
The tools of cleaning, describing and matching were developed and updated among several stages to reach a developed program that cleans and matches all data files together to be ready for manual validation
Insee
EXPERTISE FRANCE


Developing Matching Mechanisms
Links between the economic units of each partner have been developed through two internal IDs for data exchange systems among partners and not for public use.
Establishment_ID
Connects each local unit (establishment) with its correspondence in partner files
Enterprise_ID
Connects each legal unit (enterprise) with its correspondence in partner files
Insee
EXPERTISE FRANCE

Total Matching Results

Partner data was automatically matched
and manually validated. The results were as follows:

40,000
Establishments

33,000
Enterprises

Matching Matrix of Partners – West Bank

Companies									
14,041								Matched amorig all partners	MoNE
11,815									VAT
4,206									CC
3,682									Large Munici-palities
14,041	786	522	680	6,911	1,957	1,616	213	1,356	Total

Matching Matrix of Partners – West Bank

Individuals								
9,891							Matched among all partners	VAT
2,363								CC
11,355								Large Municipalities
16,843	6,920	38	4,131	1,319	3,429	4	1,002	Total

Insee EXPERTISE FRANCE

BR–Value Added for Partners

The partners provide the system with two types of data: the new units since the last update, and the updating of existing data

Helping all partners to choose which establishments or enterprises should be registered in each partner's file

Establishments and enterprises are the units that will help in linking the partners units in a smart way to allow partners knowing if they have registered / licensed all units that should be in their files

Each partner keeps its files and its own computerized system

Insee EXPERTISE FRANCE

The Road Map to Move from the Current Register to a More Mature System

To go to municipalities to increase coverage based on the establishments census 2017

Re-evaluation by linking with the establishment census 2022 and putting the final steps to exit with a register of higher quality and comprehensiveness

Summary

The business register is developed only through improved partner data.
This requires each partner to provide data to PCBS,
which in turn will improve the business register through a well-trained team

Building a mature business register needs cooperation of partners
by providing PCBS with the best quality data on regular basis

PCBS will provide a qualified team to improve the Register at the current stage

Business Register is a very important tool to build a standard unified window

Insee

EXPERTISE FRANCE

THANK YOU

Insee
Mesurer pour comprendre

EXPERTISE FRANCE