B12: Data Matching Across Agencies to Support Interoperability

- Who is this tool for? Health and social protection leaders working on data linkage projects to improve population targeting, and at the stage of implementation where they are planning their strategies for matching records across the agencies and datasets concerned.
- How was it produced? A workshop on records validation and matching was held in July 2021 to support one of the implementation case countries in the collaborative, Ghana, with this particular topic. General insights produced in the course of this meeting have been distilled below as global guidance on the key issues and approaches.

Data matching is a crucial step in achieving data sharing across agencies. It involves bringing together data from different sources and comparing it. At its most basic, you are trying to determine the probability that records on households or individuals refer to the same entity in the databases that will share data. Several terms are used for this process: (i) record or data linkage; (ii) entity resolution; (iii) object identification; and (iv) field matching. Data matching can refer to matching across datasets (the main focus here), or for a single data set can be used for deduplication.

Ideally this process should take place after building blocks that are described under other sections of this toolbox are in place or have occurred. They include:

- A basic governance framework to guide technical dimensions of interoperability – see sections on whole-of-government examples of this.
- An interoperability framework to formalize a standardized approach, and specify the political and legal context, the business processes and concepts involved in interoperability operations, and the technologies used to implement them. – see section on interoperability frameworks (Tool B3).
- A data sharing service agreement or protocol of some form between the relevant agencies. This may be high level and fairly general (e.g., Indonesia Unified DB MOU template) or more elaborated, including business rules and procedures for resolving conflicts in data from different sources. The section on data sharing agreements covers this in more detail (Tool B4).
- Data cleaning should ideally be done by each agency prior to the matching process to ensure that the data in each database is as clean as possible to reduce sources of mismatch prior to the data matching stage. The section on data cleaning covers this in detail (Tool B11).
- If not already done through reliance on common data standards across agencies/databases, then you need to standardize (or normalize) the data to ensure a consistent format for matching. This allows the mapping of the data from its given format into the format expected by the other database or application with/through which data will be shared across agencies. This includes for example value conversions or translation functions, as well as normalizing numeric values to conform to minimum and maximum values. This should ideally be done in the home agency but

can also be applied by the receiving agency software or through use of middleware (which is software that lies between an operating system and the applications running on it). Standard softwares to do rules-based data transformations are available or more complex procedures can be used where there is capacity to do so.

A threshold question to ask prior to initiating data matching across agencies is whether there is a common/unique identifier for households and/or individuals. If not, can one be created ? If a unique national ID exists and is used by both agencies, then this element is fairly straightforward. But if it is not available and incorporated in the data to be shared, there are alternative ways to establish a unique number across databases – e.g., Brazil uses "match key" variables (name; mother's name; DOB; and codes from documents); Philippines uses a probability model to assess the likelihood of matches.

In terms of preparing data for matching XML and JSON are competing methods of organizing complex data in understandable and readable format to various APIs and programming languages. XML is an extension of HTML (the web-creation language) and JSON is the use of Javascript programming language to describe documents. Both can be considered "meta-data languages" as they describe the underlying data so that you can make sense out of the datastream, and greatly facilitate data sharing. You can use both with any programming language. And they are both human and computer readable.

Both methods are ways of "tagging" fields rather than designing records with fixed fields. For example, in conventional programming a record might have fields: Last Name, then First Name, then Middle Initial, then Nickname, in a fixed order. In XML and JSON, the fields can be in any order and are tagged: for example: <First name> <Nickname> <Middle Initial> <Last Name>; and some can be omitted (say you have no nickname) and some can be duplicated (say you have two middle initials). See: https://www.imaginarycloud.com/blog/json-vs-xml/ on pros and cons of using XML vs. JSON.

After considering those prior factors, agencies move to the matching or entity resolution stage, which is the core of the matching process. This involves looking at all the information on households or individuals from both data sources and applying likelihood or probability scoring to identify matches. It relies on an entity resolution engine or middleware.

There are two initial steps which can be useful in terms of organizing data for a more efficient matching process:

- **Blocking**, which groups records in the same group or block as the basis for focusing the matching process (e.g., grouping records from the same ZIP code, geographic region, etc).

- **Indexing,** which is done to reduce the complexity resulting from the data matching process and to build an effective index structure that can potentially generate matching data. This step can involve the selection of the data column to ensure data can be compared and matched.

After first blocking and/or indexing, broadly there are two approaches to linking the data across sources, both of which rely on running algorithms on the data. There are many available softwares which can support this process. A summary and comparison of around 20 such softwares is provided at https://github.com/J535D165/data-matching-software. The comparison looks at the following features of each software: (i) application programming interface (API); (ii) graphical user interface (GUI); (ii) linking; (iii) deduplication; (iv) supervised learning; (v) unsupervised Learning and (vi) active learning. The software in the list is open source and/or freely available. The two main matching approaches are:

- **Deterministic record linkage**: this takes matching identifiers across data sets and sets a threshold level of similarity to determine matches of records across the data sources. It involves looking for exact matches in records according to the threshold. This works well for example when there is a common identifier across databases such as a national ID number.

- **Probabilistic record linkage or "fuzzy matching"**: this uses a wider set of identifiers on records and applies weights to identify matches/non-matches and the probabilities of matches. In fuzzy matching, you determine a threshold level above which two records can be considered a match and another minimum threshold below which records are considered a non-match. The appropriate thresholds for match/non-match can be predicted by record linkage software. In between those will be a group of "possible matches" which need to be dealt with to determine if they are matches or non-matches (e.g., by being human reviewed). This is the "fuzzy" ground in the matching process. Some of the algorithms used apply different underlying approaches, which include: Levenstein distance (minimum number of single-character edits - insertions, deletions or substitutions - required to change one word into the other), Jaro–Winkler distance, and regular-expression comparison.

In addition to the matching options above, there is also the possibility of what is known as **triangulation confirmation**: These algorithms examine the connections between different matching algorithms run on the data. They allow more specific classification of the data.

An important consideration in data matching exercises with personal data is **data privacy/personal data protection** provisions in the country. In countries with well-developed legal provisions for person data protection, this may impose constraints either on what data can be matched across agencies, and/or may require certain procedures for managing data in the data matching process, e.g., depersonalization of certain data.

Examples of data matching guidelines*:*

From **Australia,**

https://www.oaic.gov.au/privacy/guidance-and-advice/guidelines-on-data-matching-in-australian-government-administration.

This comes from the Australian Govt . Of particular use on probability data linking is the document available here (it is a sub-tag of the overall link toolkit.data.gov.au).

https://toolkit.data.gov.au/Data_Linking_Information_Series_Contents_page.html

From **Wales**, there is a Code of Data Matching, 2018.

https://www.audit.wales/sites/default/files-old/download_documents/code-of-data-matching-practice-english

Additional Link:

https://prateekvjoshi.com/2014/01/11/what-is-fuzzy-matching/#content