B. Issues encountered when merging data from different sources

B. Issues encountered when merging data from different sources

7.9. *Integration of different data sources.* To achieve full coverage of the international merchandise trade statistics, data compilers often have to merge and cross-check data collected from customs and non-customs sources, which is a highly complex and time-consuming activity. Merging customs and non-customs data includes adding non-customs data to the customs data and substituting non-customs data for the customs data. For the purpose of quality control and/or for the information of the users, compilers might wish to differentiate data based on customs data sources and data based on non-customs data sources.[1]

7.10. *Issues encountered when merging data from different sources*. Compilers should need to be aware that the following issues need to be addressed when merging data from different sources:

(a) Different sources may provide different data elements or levels of detail, e.g: parcel and letter post records might not contain any commodity detail; cross-border surveys might provide data only at the higher HS levels (e.g., that of HS chapters); and commodities that are difficult to classify might be allocated to a few broad categories in non-customs sources, making it difficult to merge them with the more detailed customs data (see the example of Uganda's Informal Cross Border Trade Survey below);

(b) Some transactions might be subject to simplified reporting requirements at customs;

(c) There may be conceptual differences between sources: e.g., enterprise records might contain the country of purchase and sale but not the country of origin or last known destination;

(d) There may be delays in data forwarding by some source agencies or these agencies may use different release calendars, which may lead to unsynchronized provision of data;

(e) There may be a risk of double counting due to overlaps in the information provided by different sources: e.g., between data on goods on consignment supplied by customs, and data on sales of the same goods reported by the controlling governmental agency;

(f) It may be difficult to organize data processing in an efficient manner, since source agencies may use different data submission media (hard copies, portable storage, electronic transmission, e-mail, etc.) or incompatible computer data files (the integration of different hardware and software systems is a problem in numerous cases);

(g) Data entry from certain sources (e.g., postal forms, passenger manifests) may involve the use of a disproportionate amount of time and resources;

(h) There is a need to cross-check data from complementary sources (e.g., customs and commodity boards) and to determine which sets are of greater reliability;

(i) Survey results that apply to a period longer than the reference period used for the compilation of trade statistics cannot be easily added to the customs data**;**

(j) It is not always possible to identify partner countries in detail and some rest categories will need to be used at times;

(k) The statistical value is made up of several components, some of which may not be available in some cases**;**

(l) In enterprise surveys, quantity information is frequently not collected, or cannot be provided at a level of sufficient detail.