TOSSD data architecture

TOSSD Task Force Issues Paper[1]

Ottawa, 5-6 June 2019

For discussion under agenda item 6

I. BACKGROUND

1. With the finalisation of the emerging Reporting Instructions, and the completion of the first TOSSD data survey, the Task Force is called upon to design the data architecture to be put in place for the regular data cycles. The design of this architecture will have to take several elements into account, such as:

- **The participating institutions** – institutions producing the data at the national level and the national focal points; multilateral organisations and their focal points; the custodian agency(ies) and the Secretariat; the governance structure – **and the relationships between them**.

- **The data flow and processes** i.e. how the data are collected and by whom, how they are transmitted, verified, edited, stored and disseminated. This includes the connection between the TOSSD data cycle and the data being produced in other contexts, such as CRS and IATI, and the eventual feedback loops. As and when a co-custodian agency is identified, the data flow and relevant processes may need to be adjusted.

- **The desired features**, such as process automation, scalability, modularity, portability, speed, usability, data confidentiality and security.

- **The technology choices**, meaning different kinds of IT solutions to support the data architecture, such as centralised, cloud or distributed databases. The choice of the system will have to take into account the relationships between institutions, the data flow and the desired features, but also issues like set-up and operating costs, robustness and flexibility.

2. This note discusses the main elements necessary to design the data architecture of TOSSD and the main approaches available for the design of the database.

II. Participating institutions

3. The TOSSD statistical framework has, so far, been designed with a hierarchical structure with three types of institutions:

- **The Reporters** are the focal points of the countries or the multilateral institutions participating in TOSSD. The focal points are responsible for collating, verifying and transmitting the data on TOSSD-eligible activities within the country or the participating institution. The focal points are responsible for feeding the data into the TOSSD statistical framework and for responding to any feedback from the Secretariat (custodian agency). They shall produce data that comply

[1] Drafted by Giorgio Gualberti (Giorgio.Gualberti@oecd.org), Valerie Thielemans (Valerie.Thielemans@oecd.org) and Julia Benn (Julia.Benn@oecd.org).

with the agreed Reporting Instructions and they are ultimately responsible for verifying and signing off the data.

- **The Secretariat** collects the data, runs quality checks, assures that focal points follow the agreed rules and procedures, sends feedback to the focal points, manages the IT architecture and publishes the data. It also organises meetings on TOSSD and serves all the functions that the governance body will assign.
- **The Governance body** takes the decisions about the TOSSD rules and procedures and oversees that the Secretariat and the Reporters conform to them. This role is *ad interim* performed by the Task Force.

4. A fourth type of institutions, while not *directly* participating in TOSSD, have a significant role in the system. These are the national or subnational institutions that finance or implement the TOSSD activities, or the national or regional offices of the multilateral institutions. These institutions may be the original producers of the data that are later transmitted to the focal points and ultimately reach the TOSSD Secretariat. In some cases, these institutions also publish their data independently in IATI.

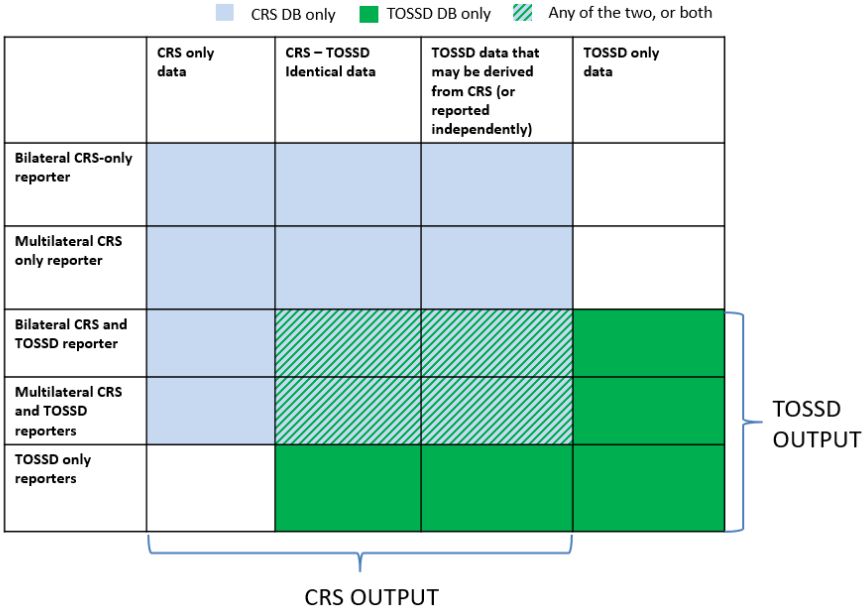
III. The data flow and processes

The data ecosystem

5. The TOSSD dataset, as designed for the data survey, is composed of both data that are generated specifically for TOSSD, and information that might be already produced for other reporting frameworks. As such, the TOSSD data cycle is not independent but lives in a data ecosystem with strong interlinkages.

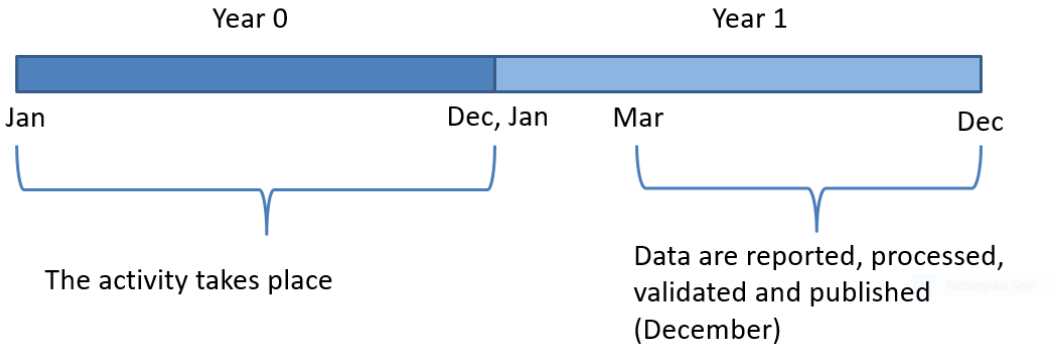
6. The linkages between the TOSSD and the CRS systems are significant. There are around 20 data fields in the TOSSD data format that can be derived from CRS fields, so a significant data overlap exists (Figure 1). A majority of CRS data reporters have expressed interest in participating in TOSSD.

Figure 1 – Overlaps between CRS and TOSSD data providers and outputs (DB=Database)



7. The CRS data cycle is well-established and reliable. National focal points report every year to the OECD activities that were undertaken in the previous year. The Secretariat checks the data, sends feedback to the data providers, performs all the data treatment procedures and then publishes the new dataset at the end of the year (Figure 2). A delay between 12 to 24 months could exist from the moment an activity occurs until it is published online by the OECD in the CRS. The CRS dataset is updated on a quarterly basis but the data cycle remains yearly and the full update for all institutions is completed at the end of the fourth quarter.

Figure 2- CRS data cycle timeline



8. The overlap between the data produced for CRS and for TOSSD implies that a significant part of the TOSSD data will be available – at the latest – at the end of the fourth quarter, at same date at which the complete CRS data for the previous year are published.[2] However, developing countries might be interested in having access to this information – or at least part of it – earlier. The deadline for reporting to the CRS by bilateral providers is the 15th of July regarding data for the previous year. Therefore, an earlier publication will only be possible if the data providers are able to anticipate their submission date and/or if a faster treatment of the data is possible. Capacity constraints to report data and the need to maintain the same quality control mechanisms might prevent earlier release of detailed TOSSD information.

9. A large number of institutions publish information on their development co-operation activities using the IATI standard. IATI is primarily a tool of data standardisation and exchange, but not for statistical purposes. The IATI standard allows immediate data publication by the participating institutions that retain full ownership of the data. The data are stored in a decentralised form and referenced in a common IATI-registry.

10. The IATI standard includes a set of rules defining what kind of information can be published (the list of fields / subfields, the vocabularies allowed for each of them) and what data exchange format can be used (a customised XML file). Several IATI fields / subfields and vocabularies originate from the CRS fields and code lists (e.g. the purpose codes, modalities) and are therefore directly linkable to the TOSSD fields. In principle, it is possible to use data published in the IATI format to pre-fill the TOSSD compatible fields, with the main limiting factor being uncertainty about the quality and completeness of the underlying data.

[2] Provided the TOSSD eligibility of the activities, in particular whether they comply with the criterion of sustainable development, has been verified and the data items specific to TOSSD (e.g. salary costs) have been compiled.

11. The Secretariat has tested this procedure with the IATI submission for two multilateral institutions and found that although it is technically possible, the results are dependent on the completeness of the IATI submission. In a first case, the IATI XML files analysed were complete and no major impediment was found in the process. The data were used to pre-fill the TOSSD data survey that was then submitted to the focal point to be validated and integrated with the TOSSD specific information. In a second case, the conversion was still possible, but the file was lacking some TOSSD mandatory fields (such as the modality and the financial instrument) and needed to be sent to the focal point for completion.

12. Besides data completeness, the quality of the IATI data varies greatly from institution to institution, due to the lack of a centralised control mechanism. Data quality issues, that include data gaps, possible double reporting of the same project by more than one institution, or other erroneous or non-consistent reporting, will need to be scrutinized in detail before IATI data can feed in TOSSD.

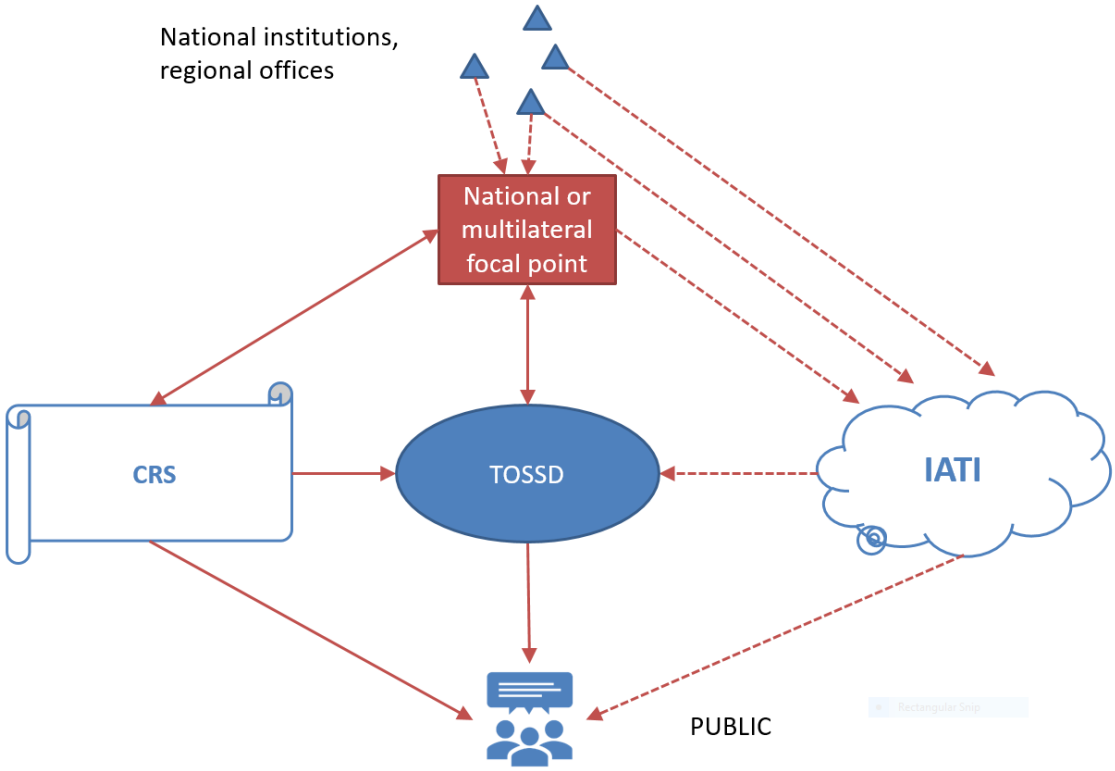
13. The use of IATI data to pre-fill part of the TOSSD data fields could be particularly useful for multilateral institutions, given that the scope of multilateral activities reported in TOSSD is broader than in the CRS. For any institution that publishes comprehensive data in IATI format, such a mechanism would lower the reporting burden to fill the statistical gap.

14. The data available in IATI format could also be considered for another reason: providing provisional data on TOSSD activities at an earlier date than the one possible with the regular data cycle. The Secretariat held a series of meetings with data providers that participate in both CRS and IATI exercises to understand how they manage the double submissions internally, and how they understand the practicality of this idea. Initial results show a great variety of approaches. Some reporters have centralised systems that are used to source the data for both CRS and IATI, and the resulting submissions are mostly aligned. Some others have independent systems for IATI and CRS submissions, and in some cases the CRS and IATI submission differs considerably, for example when not all the subnational institutions that are included in the CRS submission also report to IATI. Further research is however needed in this area, if the Task Force expresses interest.

The data cycle

15. To minimise the reporting burden, on the occasion of the TOSSD data survey, the Secretariat provided the data reporters with excel sheets pre-filled with data harvested from CRS (and in two cases IATI) submissions, when available. The Secretariat then asked the reporting institutions to provide the additional TOSSD-only elements and to validate the data. A similar system could be put in place for the regular data collection (Figure 3).

Figure 3 – Possible data flows and interactions between TOSSD, CRS and IATI



Note: This Figure does not take into account other possible actors involved in the collection of TOSSD data such as regional data hubs.

16. The TOSSD data flow will differ depending on the type of institution involved and the availability of the data. For example:

- **Bilateral institutions that report to CRS:** these institutions could be asked to fill additional "TOSSD-only" data fields, and provide to the OECD the additional "TOSSD-only" records alongside their regular CRS data. These additional data could then be treated in parallel with CRS data. However, if this solution were to be adopted, it must not lead to delays in CRS reporting for which the deadline is 15 July each year.

- **Multilateral institutions that report to the CRS or IATI.** The data flow described above can apply equally to multilateral institutions reporting to the CRS. In some cases, however, it might be advantageous to start with IATI submission, if the comprehensiveness and the quality of the data are deemed sufficient. The data should then be harmonised with the TOSSD format and sent to the focal point for completion and validation.

- **Other data providers** that do not participate in either CRS or IATI data collection would send a complete dataset directly to the Secretariat. The Secretariat will provide feedback and ask for validation.

17. The three types of data flow would undergo the same quality controls and, once validated by both the Secretariat and the data provider, be published in the TOSSD system.

IV. Desired features

18. A series of features should be taken into consideration for the TOSSD data architecture. These can include the following:

- **Distributed Access.** Data providers would be able to get direct access to the data that they submit to the TOSSD database, and directly add/edit data as they see fit without necessarily passing through the Secretariat. The Secretariat and the data providers would need to agree on the revisions, assuring appropriate quality control and feedback mechanisms. A historical log of all changes would be desirable.

- **Application Program Interface (API).** The data should be able to flow directly from one system to the other without human intervention. Data submission by email should be actively discouraged, and not allowed at all in case of confidential data. Other websites and programmes should be able to interface directly with the database and access all non-confidential data. The Secretariat could assist data providers with more limited technical capabilities in the data submission.

- **Process automation.** The system should be able to run internal checks to verify the integrity and quality of the data, and flag activities for revision. The system can use machine learning to check the sector codes or the SDGs and help identifying misattributions.

- **Confidentiality**. All TOSSD data will be made publicly available at the activity level. Reporting Instructions invite providers to filter out upstream any information linked to TOSSD activities subject to confidentiality regimes. This might apply in particular to some operations involving private finance mobilised.

- **Scalability and modularity.** The system should be flexible enough to be able to include additional modules (such as satellite indicators, or links to external data) and to scale-up its capacity to store and retrieve data without bottlenecks.

- **Portability.** The TOSSD database and ancillary software (such as the API, quality checks algorithms etc.) should be portable, i.e. it should be possible to move it from one server or one system to another.

- **Speed.** The system will need to store large number of records (possibly more than 250 000 a year) and retrieve information quickly and reliably.

- **Usability.** The system would need to be used by many people in a large number of different institutions. While the inner workings can be complex, the interface with both the data providers and the users would need to be clearly accessible. Simple operations like uploading data, verifying or editing data, selecting data for download should not need a steep learning curve.

- **Future-proofing.** The design and technology choices adopted shall be able to stand for the foreseeable future.

19. The features listed above are only indicative, and serve the purpose of starting a discussion among Task Force members on the desired characteristics of the system. Technical and human

capabilities, as well as the cost and complexity of the different solutions shall also play a role in the final design choices.

V. Database designs

20. The Secretariat started a series of bilateral meetings with database experts and practitioners to identify the possible design to be adopted to support the desired features. Three types of design have been taken into consideration.

    - **Internal corporate database**. This type of database runs on the internal servers of the Secretariat. It is the interim solution adopted for storing the results of the TOSSD data survey. The advantages of this solution include the possibility of building the database in-house, taking advantage of the existing infrastructure and staff capacities with limited additional costs. It also has the advantage of allowing an easy transfer of the overlapping records from CRS to TOSSD. On the other hand, the interim corporate database put in place has less flexibility in terms of advanced features, such as APIs.

    - **External or cloud database**. A database can be run on external or cloud servers. This solution allows third parties to directly submit and edit their data, permitting faster publication times, and could be easily scalable to match the needed capacity. Technology providers could offer a wide range of solutions to satisfy the needs as they appear.

    - **Distributed ledger technology (DLT)** databases are spread across several nodes (servers) in a peer to peer network. These technologies, that include blockchain, are being increasingly adopted in particular for their characteristic of robustness, security and capacity to keep an historic track of decentralised transactions. DLTs databases can adopt a wide array of underlying solutions and they are certainly more complex to set-up than traditional corporate or cloud databases, but offer some unique capabilities.

21. Each type of design has advantages and disadvantages. The choice will depend on considerations on which technologies and design is a better fit for accommodating the TOSSD data flow and providing the desired features for the system. The choice will also depend on the available financial resources for implementing and operating the TOSSD system. The Secretariat is exploring all options available and welcomes inputs and the sharing of experiences from Task Force members.

Issues for discussion

1. **Do Task Force members have comments on section II, which describes the participating institutions and their roles in the TOSSD data architecture?**
2. **Do Task Force members have comments on the initial analysis of the TOSSD data flow described in section III?**
3. **Do Task Force members have comments on the desired features of the TOSSD data architecture listed in section IV?**
4. **Do Task Force members have views or experiences to share on the database design choices described in section V?**