Developing a Logical Information Model for the Common Statistical Production Architecture

March 2105

I. Current situation and issues

1. The Generic Statistical Information Model (GSIM) provides a common language that increases the ability to compare information within and between statistical organizations. It describes at the conceptual level the information the statistical production process consumes and produces.

2. GSIM describes the information objects relevant to statistical production in a way that helps people communicate – as good conceptual information models should. However, it does not provide enough detail for implementation – that is; it does not help CSPA services communicate with each other.

3. The primary interest for the designer and the builder of a CSPA service is likely to be the physical specification of the information that will flow into and out of the service. At the current time choices for physical specifications might include:

- SDMX-ML (Version 2.1) (SDMX represented in XML),
- SDMX-JSON (SDMX represented in JSON)
- DDI 3.2 XML
- CSV

4. These choices often prove confusing for service designers and builders. There is a lack of consistent guidance to help designers and builders determine which choice is likely to be most appropriate for the purpose of the specific service they are developing.

5. In addition, not all choices are identical in terms of logical definition of information. For example, while a SDMX-ML representation of a code list and a DDI 3.2 representation of a code list are similar in many regards, they are not the same and are not fully interchangeable/ interoperable. If one service accepts an SDMX-ML codelist as an input and another service accepts a DDI 3.2 codelist as an input, the user may need to redefine or restructure their codelist between using the first service and using the second service. In other words, information would not flow efficiently between services.

6. The definition of the CSPA Logical Information Model (LIM) and the specification of physical representations based on this logical model provides the way forward in addressing "the missing link" for translating agreed GSIM concepts to consistent, standards aligned, physical inputs and outputs for CSPA services.

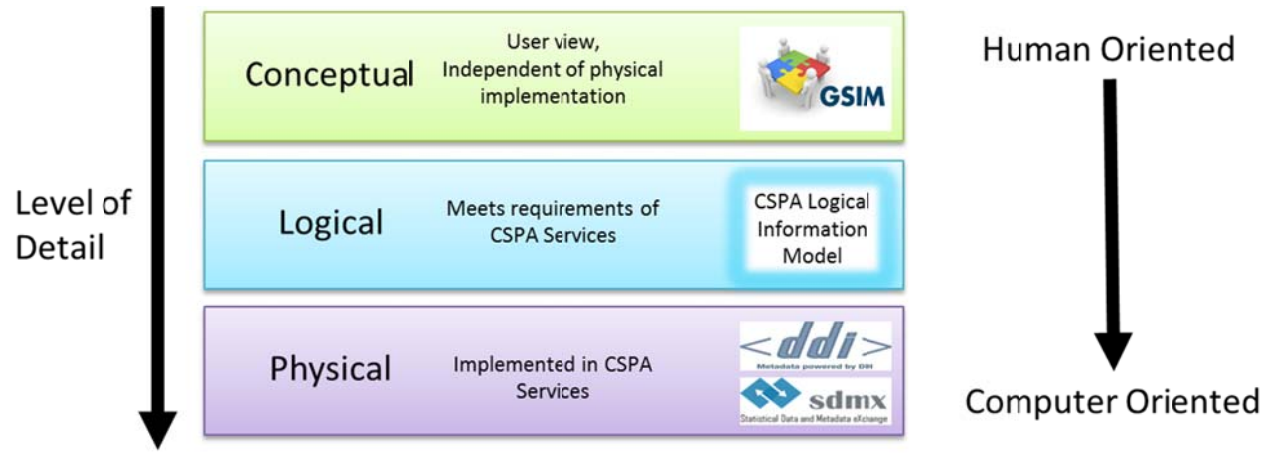

Figure 1. Conceptual, logical and physical models

II. CSPA Logical Information Model

7. In order for interoperability and reuse to be supported in practice when applying CSPA, the statistical modernization community needs to do more than align conceptual designs using common frameworks.

8. For Service Specification, we need to describe information objects and the precise logical relationships between them in a manner which is consistent with GSIM. Establishing a consistent set of objects and attributes independent of the terminology used in existing standards such as SDMX and DDI requires the development of a CSPA Logical Information Model (LIM).

9. Standards such as SDMX and DDI offer examples of such detailed logical models. They identify and use many more attributes than are defined within GSIM. As these standards and practices have evolved independently, the objects and attributes they use are similar but not consistent and can be implemented in many different ways.

10. The CSPA LIM will:

- complement, and be founded upon, the existing GSIM conceptual model;
- support consistent use of SDMX, DDI and other implementation standards in reusable CPSA services;
- make it easier for organisations that do not use SDMX or DDI to implement reusable CSPA services, and
- be a useful way for users to precisely document the inputs and outputs of their business processes.

11. The first stage when specifying the information that will flow in and out of a service will be to model the required information within the LIM. Logical modelling will be undertaken in a manner which is consistent with GSIM, the agreed conceptual model for statistical information used by CSPA services. It will be quickly and readily possible to identify which GSIM

information objects (defined at the conceptual level) are relevant to the inputs to, and outputs from, the new service which is being designed.

12. The next step for logical modelling will be identifying which existing standards (if any) are relevant to the GSIM (conceptual) information objects which are in scope of the particular service. Logical modelling for CSPA will align to the maximum practical extent with the logical models associated with the candidate standards. In cases where complete alignment with existing standards is not practical, the usual decision will be for the LIM to align with one or other of the choices on a "best fit" basis.

13. While the CSPA Logical Information Model will act as a bridge at the logical level between different standards, it will be possible to make physical representations of information objects in CSPA in formats defined by existing standards.

14. Depending on what information is being represented in practice, DDI and SDMX are currently expected to provide the primary basis for the physical representation of statistical information (e.g. data and metadata) in CSPA.

15. The figure below shows the CSPA LIM and its possible physical representations. LIM will make use of a number of standard logical models, e.g. DDI, SDMX, and others.

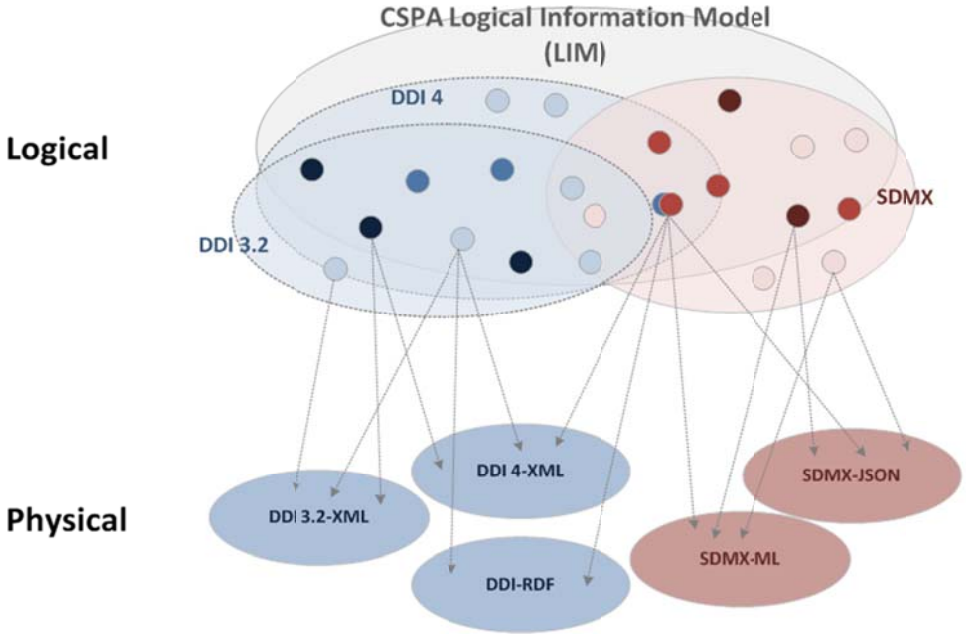

Figure 2. CSPA LIM and its possible physical representations

16. There is an overlap between standard logical models which means there are cases where it would be feasible for a particular LIM object to be represented by different standards. Annex 1 - under the sub-heading "Case 1" - describes how it is proposed to address cases of overlap.

17. There are also some areas in the standards that are outside LIM. Those areas consist of objects around which there is no clear usage agreement among practitioners: people in different

domains represent the same information by means of different logical objects. Those ambiguous areas will be reduced over time as we include more objects in LIM and we disambiguate their usage for data exchange among CSPA services.

18. Further detail on how physical implementation guidance will be provided to designers and builders is in Annex 1.

III. How LIM will be developed - an incremental approach

19. The development of LIM and its physical representation will be undertaken incrementally based on business requirements. GSIM includes a wide variety of objects and not all of them are required by services with the same urgency. In addition, the effort required to produce an all-encompassing logical model in a short time might prove to be prohibitively expensive for the CSPA project and the participating statistical organizations.

20. Requirements for the work will have four different drivers, as described below. It is expected the first driver will be assigned highest priority. There will be strong benefits, however, if sufficient resources are available to undertake modelling which also addresses the other three drivers.

CSPA services development roadmap

21. At any given point in time there will be CSPA services that require specific information objects for data exchange. Priorities and timelines of the services, together with dependencies among them, will be taken into account to develop an object modeling plan. Providing the logical and physical models for objects in the design and implementation phases is critical.

Statistical organizations internal service development roadmap

22. Similar to the previous point but internal to the participating organizations. Services for internal consumption might benefit from consultation with the CSPA community to determine the feasibility of implementing them with standard GSIM objects and potentially sharing them with other agencies. This discussion should also inform the object modeling plan.

Reusability factor

23. Another criterion to decide where a given object fits in the roadmap is to what extent it is expected to be shared across services. Objects belonging to the first and last phases of the GSBPM are initial candidates, but further analysis is required to find key reusable objects across all GSBPM sub-processes.

Coverage provided by existing standards

24. Not all conceivable LIM objects will have an equivalent logical object in one of CSPA's preferred standards, e.g. DDI, SDMX, etc. We will start with those LIM objects that can be

represented using at least one of the standards and integrate other objects as required (e.g. process metadata). Eventually, as standards evolve, more objects will be covered.

IV. Logistical concerns

25. The process to implement the roadmap will need to address several logistical concerns. These might cover such areas as:

Resources

26. The work to be carried out will require investment by interested organizations, in terms of money and people.

Modelling environment/infrastructure

27. The modelling process for the CSPA logical model would need a suitable collaborative work environment. The platform adopted for previous development of the GSIM was Atlassian Confluence with a Gliffy diagramming plug-in. However, this type of tool does not enforce any sort of integrity or consistency checking of the sort which would be found in other modelling tools such as Enterprise Architect (EA). The use of such a tool as EA, however, will incur expense for licences. This would require further discussion to weigh the costs and benefits.

Annex I: Developing the LIM in stages.

28. There will be three stages when providing guidance to service designers and builders.

Stage 1

29. In future, the first stage when specifying the information that will flow in and out of a service will be to model the required information within the LIM.

30. As more logical modelling within LIM is completed, there will be more cases where the required information has already been modelled. In those situations it will simply be a case of confirming whether the existing modelling fully meets the needs of the new service or whether one or more minor extensions to the logical model are required to more fully address the information needs of the new service which is being designed.

31. Logical modelling will be undertaken in a manner which is consistent with GSIM, the agreed conceptual model for statistical information used by CSPA services. It will be quickly and readily possible to identify which GSIM information objects (defined at the conceptual level) are relevant to the inputs to, and outputs from, the new service which is being designed.

32. The next step for logical modelling will be identifying which existing standards (if any) are relevant to the GSIM (conceptual) information objects which are in scope. One of three cases will arise.

<u>Case 1: More than one standard is a plausible candidate for representing the GSIM object</u>

33. Some GSIM information objects could be implemented using more than one standard. For example, Codelists and Data Structures can be represented in both SDMX and DDI.

34. In this case the logical models associated with the candidate standards will be analysed. Some standards do not have a separately defined logical model. In such cases, however, a logical model can be "reverse engineered" from the specification - such as an XML schema - within the standard which defines the physical representation.

35. Logical modelling for CSPA will align to the maximum practical extent with the logical models associated with the candidate standards. In cases where complete alignment with existing standards is not practical the usual decision will be for the LIM to align with one or other of the choices on a "best fit" basis.

36. During Stage 1, the recommended logical modelling within CSPA will not yet have been reviewed and confirmed by the broader community of CSPA stakeholders. A service designer might choose to implement the input or output using another standard. For example, if logical modelling from SDMX is recommended for a particular GSIM object then the service designer might still choose to implement the input using DDI.

37. However, if a service designer chooses to do that, mapping from LIM to the logical model of the other standard becomes the responsibility of the service designer/builder. The service investor incurs the risk that the recommended logical modelling will not subsequently (in Stage 2) be confirmed by the CSPA community and not become mandated for CSPA services in future. Should that happen the service investor would have the option of:

- remediating the service to use the mandated logical model for inputs and outputs, or
- having the service catalogued as not using CSPA compliant inputs and outputs.

Case 2: Only one standard is a plausible candidate for representing the GSIM object.

38. This is likely to be the most common case and the most straightforward case. The logical model associated with the standard will be checked for basic fitness for purpose. If it passes that check then it will provide the basis for modelling within LIM.

Case 3: No existing standard is identified as a plausible candidate for representing the GSIM object

39. This case will require additional research and analysis by the modelling team. An existing standard may provide a starting point for the modelling, but significant elements of the modelling may also be specific to CSPA.

40. Such modelling may carry additional uncertainty or risk. The model will not previously have been trialed through implementation via application of an existing standard.

41. An agile approach to LIM may allow only the subset of information required by the service to be modelled initially. The new entity in the logical model might then have additional attributes and relationships added later once information requirements from future, not yet specified, services which related to the same GSIM information object have become apparent.

42. Where the GSIM object was not accommodated in an existing standard, but would be a sensible addition to that standard, the Executive Board may suggest to the relevant standards body that a future edition of the standard incorporate equivalent modelling to LIM. Application of the modelling within CSPA, however, would take place regardless of the ultimate decision of that standards body on whether or not to incorporate the feature.

43. Stage 1 will result in advice to the service designer on the recommended modelling of the inputs and outputs.

44. In Case 1 and Case 2, the recommended logical model will reference the relevant representation standard. For example, if logical modelling based on SDMX is selected then implementation using SDMX-ML will be an obvious choice for physical representation. Other representations based on the SDMX Information Model (eg SDMX JSON) may also be valid choices.

45. In Case 3 a very simple "CSPA" schema would be provided (eg in XML). Physical implementation would be based on that schema.

Stage 2

46. Stage 2 relates to confirming the recommended logical modelling and mandating it as standard within CSPA. While, depending on existing resources and competing priorities, Stage 2 need not occur immediately after Stage 1, it is advantageous if it does.

47. It is important to recognise that the initial advice to the service designer and builder is not dependent on Stage 2.

48. Stage 2 would allow broader review of the recommended logical modelling within the "CSPA community". Comments may be received from, for example,

- statistical organizations that have their own logical models and/or existing business processes and services related to the GSIM information object(s) in question
- standards bodies (eg for SDMX and DDI)

49. Stage 2 will add assurance that the recommended modelling provides a sound way forward. Unless feedback from the broader community identifies a critical oversight in the recommendation provided at the end of Stage 1, the specifications provided to the service designer previously will not be changed in a fundamental manner that would revise the information interface for the service. It is more likely that the advice and modelling will be clarified. Additional attributes may be identified through consultation which should be added to the model - but only if there is a common, practical, well defined use case for adding them.

50. The aim is to ensure that while the modelling may need to be extended to meet additional needs from new services designed in future, the logical model should not need to be fundamentally restructured in future. This means existing services which specify physical inputs and outputs based on the logical model will not be impacted if, and when, the logical modelling is updated (extended) in future to support the needs of new services.

51. At the conclusion of Stage 2, the logical modelling will become "CSPA mandated". Services which did not use the CSPA logical modelling (e.g. which provided an information interface based on DDI variant of the modelling where SDMX provided the basis for the modelling in CSPA) would not be CSPA compliant in regard to their inputs and outputs.

52. In this example, an agency which "preferred" DDI over SDMX might choose in their own environment to:

- translate from DDI to SDMX when providing inputs to the service, and
- translate from SDMX to DDI when accepting outputs from the service

53. Such translation would be the responsibility of the agency. While "adaptors" between common standards such as SDMX and DDI could be developed and shared in future to assist

with this, CSPA services themselves would only accept input in a physical format which was fully consistent with the logical model.

Stage 3

54. As described under *Current Situation and Issues*, there can exist multiple physical formats (eg SDMX-ML vs SDMX JSON) which are capable of expressing the same logical model (the SDMX Information Model for this example). Where multiple, logically equivalent, physical implementation choices exist then Stage 1 and Stage 2 would place no restriction on which physical implementation a service designer or service builder selects.

55. Stage 3 could, if the CSPA community decides to proceed this far, specify which physical implementation must be used where multiple logically equivalent physical implementations would otherwise be possible.

56. The main advantages of proceeding to Stage 3 arise from consistency in regard to physical implementation. This should make it more straightforward to implement multiple CSPA services within an agency.

57. Stage 3 avoids the possibility that designers of two different CSPA services may implement logically equivalent information inputs and outputs on a different physical basis. Should different physical bases be used by different services:

- the implementer of the services within a statistical agency will need to do more translation/preparation of information as inputs or outputs from the services
- the implementer of the services within a statistical agency may need to become familiar with more physical formats (eg XML vs JSON) than would otherwise be the case

58. Compared with the translating between "similar but different" logical implementations (eg SDMX Code List vs DDI Code List), the cost and complexity of translating between logically equivalent physical implementations is low. It is, nevertheless, unlikely to be considered "trivial", especially by the implementer asked to do it.

59. Costs and disadvantages of proceeding to Stage 3 might include:

- The builder of a potential CSPA Service may be familiar with, and/or prefer, an alternative physical rendering. This may be a barrier to the investor deciding to develop the service as "CSPA compliant" rather than simply developing the service for local purposes.
- Situations could arise where a number of inputs to a particular service will be in a particular format (eg JSON) and it would make the service simpler and more efficient if another input was in the same format. Under Stage 3, however, a different format may have been mandated for the final input.

60. A decision is not required immediately on whether to proceed to Stage 3 or not. If it is decided to proceed to Stage 3 after many CSPA services have already been made available, however, the implementation cost for moving to Stage 3 will be higher.