Statistical Data and Metadata eXchange

SDMX STANDARDS: SECTION 1

FRAMEWORK FOR SDMX TECHNICAL STANDARDS

VERSION 2.1

April 2011

Contents

1 Introduction

The Statistical Data and Metadata Exchange (SDMX) initiative (http://www.sdmx.org) sets standards that can facilitate the exchange of statistical data and metadata using modern information technology, with an emphasis on aggregated data.

There are several sections to the SDMX Technical Specification:

1. SDMX Framework Document – this document. The purpose of this document is to introduce SDMX and its scope. This document will be revised in due course to include the conformance requirements.

2. The SDMX Information Model - the information model on which syntax-specific implementations described in the other sections are based. This is intended for technicians wishing to understand the complete scope of the technical standards in a syntax-neutral form. It includes as an annex a tutorial on UML (Unified Modelling Language). This document is not normative.

3. SDMX-EDI - the UN/EDIFACT format for exchange of SDMX-structured data and metadata. This document contains normative sections describing the use of the UN/EDIFACT syntax in SDMX messages. This document has normative sections.

4. SDMX-ML - the XML format for the exchange of SDMX-structured data and metadata. This document has normative sections describing the use of the XML syntax in SDMX messages, and is accompanied by a set of normative XML schemas and non-normative sample XML document instances.

5. The SDMX Registry Specification provides for a central registry of information about available data and reference metadata, and for a repository containing structural metadata and provisioning information. This specification defines the basic services offered by the SDMX Registry: registration of data and metadata; querying for data and metadata; and subscription/notification regarding updates to the registry. This document has normative sections.

6. The SDMX Technical Notes – this is a guide to help those who wish to use the SDMX specifications. It includes notes on the expressive differences of the various messages and syntaxes; versioning; maintenance agencies; the SDMX Registry. This document is not normative.

7. Web Services Guidelines – this is a guide for those who wish to implement SDMX using web-services technologies. It places an emphasis on those aspects of web-services technologies (including, but not requiring, an SDMX-conformant registry) which will work regardless of the development environment or platform used to create the web services. This document contains normative sections.

2 Changes from Previous Version

The 2.0 version of this standard represented a significant increase in scope, and also provided more complete support in those areas covered in the version 1.0

specification. Version 2.0 of this standard is backward-compatible with version 1.0, so that existing implementations can be easily migrated to conformance with version 2.0.

The 2.1 version of this standard represents a set of changes resulting from several years of implementation experience with the 2.0 standard. The changes do not represent a major increase in scope or functionality, but do correct some bugs, and add functionalities in some cases. Major changes in SDMX-ML include a much stronger alignment of the XML Schemas with the Information Model, to emphasize inheritance and object-oriented features, and increased precision and flexibility in the attachment of metadata reports to specific objects in the SDMX Information Model.

Note that the idea of backward-compatibility in the standards is based on the information model. In both releases, some non-backward-compatible changes have been made to the SDMX-ML formats. The same set of information required to use version 1.0 of the specification will permit the use of the same features in the version 2.0 specifications, however. Thus, a Data Structure Definition is easily translated from version 1.0 to version 2.0, without requiring any new information regarding structures, etc. There have been no changes to the SDMX-EDI format.

The major changes from 1.0 to 2.0 can be briefly summarized:

- **Reference Metadata**: In addition to describing and specifying data structures and formats (along with related structural metadata), the version 2.0 specification also provides for the exchange of metadata which is distinct from the structural metadata in the 1.0 version. This category includes “reference” metadata (regarding data quality, methodology, and similar types – it can be configured by the user to include whatever concepts require reporting); metadata related to data provisioning (release calendar information, description of the data and metadata provided, etc.); and metadata relevant to the exchange of categorization schemes.
- **SDMX Registry**: Provision is made in the 2.0 standard for standard communication with registry services, to support a data-sharing model of statistical exchange. These services include registration of data and metadata, querying of registered data and metadata, and subscription/notification.
- **Structural Metadata**: The support for exchange of statistical data and related structural metadata has been expanded. Some support is provided for qualitative data; data cube structures are described; hierarchical code lists are supported; relationships between data structures can be expressed, providing support for extensibility of data structures; and the description of functional dependencies within cubes are supported.

The major changes from 2.0 to 2.1 can be briefly summarized:

- **Web-Services-Oriented Changes:** Several organizations have been implementing web services applications using SDMX, and these implementations have resulted in several changes to the specifications. Because the nature of SDMX web services could not be anticipated at the time of the original drafting of the specifications, the web services guidelines have been completely re-developed.

- **Presentational Changes:** Much work has gone into using various technologies for the visualization of SDMX data and metadata, and some changes have been proposed as a result, to better leverage this graphical visualization. These changes are largely to leverage the Cross-domain Concepts of the Content Oriented Guidelines.
- **Consistency Issues:** There have been some areas where the draft specifications were inconsistent in minor ways, and these have been addressed.
- **Clarifications in Documentation:** In some cases it has been identified that the documentation of specific fields within the standard needed clarification and elaboration, and these issues have been addressed.
- **Optimization for XML Technologies:** Implementation has shown that it is possible to better organize the XML schemas for use within common technology development tools which work with XML. These changes are primarily focused on leveraging the object-oriented features of W3C XML Schema to allow for easier processing of SDMX data and metadata.
- **Consistency between the SDMX-ML and the SDMX Information Model:** Certain aspects of the XML schemas and UML model have been more closely aligned, to allow for easier comprehension of the SDMX model.
- **Technical Bugs:** Some minor technical bugs have been identified in the registry interfaces and elsewhere. These bugs have been addressed.
- **Support for Non-Time-Series Data in the Generic Format:** One area which has been extended is the ability to express non-time-series data as part of the generic data message.
- **Simplification of the data structure definition - specific message types:** Both time series (version 2.0 Compact) and non-time series data sets (version 2.0 Cross Sectional) use the same underlying structure for a structure-specific formatted message, which is specific to the Data Structure Definition of the data set.
- **Simplification and better support for the metadata structure:** New use cases have been reported and these are now supported by a re-modelled metadata structure definition.
- **Support for partial item schemes such as a code list:** The concept of a partial (sub-set) item scheme such as a partial code list for use in exchange scenarios has been introduced**.**

3 Processes and Business Scope

3.1 Process Patterns

SDMX identifies three basic process patterns regarding the exchange of statistical data and metadata. These can be described as follows:

1. *Bilateral exchange:* All aspects of the exchange process are agreed between counterparties, including the mechanism for exchange of data and metadata, the formats, the frequency or schedule, and the mode used for communications regarding the exchange. This is perhaps the most common process pattern.

2. *Gateway exchange:* Gateway exchanges are an organized set of bilateral exchanges, in which several data and metadata collecting organizations or

individuals agree to exchange the collected information with each other in a single, known format, and according to a single, known process. This pattern has the effect of reducing the burden of managing multiple bilateral exchanges (in data and metadata collection) across the sharing organizations/individuals. This is also a very common process pattern in the statistical area, where communities of institutions agree on ways to gain efficiencies within the scope of their collective responsibilities.

3. *Data-sharing exchange:* Open, freely available data formats and process patterns are known and standard. Thus, any organization or individual can use any counterparty's data and metadata (assuming they are permitted access to it). This model requires no bilateral agreement, but only requires that data and metadata providers and consumers adhere to the standards.

This document specifies the SDMX standards designed to facilitate exchanges based on any of these process patterns, and shows how SDMX offers advantages in all cases. It is possible to agree bilaterally to use a standard format (such as SDMX-EDI or SDMX-ML); it is possible for data senders in a gateway process to use a standard format for data exchange with each other, or with any data providers who agree to do so; it is possible to agree to use the full set of SDMX standards to support a common data-sharing process of exchange, whether based on an SDMX-conformant registry or some other architecture.

The standards specified here specifically support a data-sharing process based on the use of central registry services. Registry services provide visibility into the data and metadata existing within the community, and support the access and use of this data and metadata by providing a set of triggers for automated processing. The data or metadata itself is not stored in a central registry – these services merely provide a useful set of metadata about the data (and additional metadata) in a known location, so that users/applications can easily locate and obtain whatever data and/or metadata is registered. The use of standards for all data, metadata, and the registry services themselves is ubiquitous, permitting a high level of automation within a data-sharing community.

It should be pointed out that these different process models are not mutually exclusive – a single system capable of expressing data and metadata in SDMX-conformant formats could support all three scenarios. Different standards may be applicable to different processes (for example, many registry services interfaces are used only in a data-sharing scenario) but all have a common basis in a shared information model.

In addition to looking at collection and reporting, it is also important to consider the dissemination of data. Data and metadata – no matter how they are exchanged between counterparties in the process of their development and creation – are all eventually supplied to an end user of some type. Often, this is through specific applications inside of institutions. But more and more frequently, data and metadata are also published on websites in various formats. The dissemination of data and its accompanying metadata on the web is a focus of the SDMX standards. Standards for statistical data and metadata allow improvements in the publication of data – it becomes more easily possible to process a standard format once the data is obtained, and the data and metadata are linked together, making the comprehension and further processing of the data easier.

In discussions of statistical data, there are many aspects of its dissemination which impact data quality: data discovery, ease of use, and timeliness. SDMX standards provide support for all of these aspects of data dissemination. Standard data formats promote ease of use, and provide links to relevant metadata. The concept of registry services means that data and metadata can more easily be discovered. Timeliness is improved throughout the data lifecycle by increases in efficiency, promoted through the availability of metadata and ease of use.

It is important to note that SDMX is primarily focused on the *exchange* and *dissemination* of statistical data and metadata. There may also be many uses for the standard model and formats specified here in the context of internal processing of data that are not concerned with the exchange between organizations and users, however. It is felt that a clear, standard formatting of data and metadata for the purposes of exchange and dissemination can also facilitate internal processing by organizations and users, but this is not the focus of the specification.

3.2 SDMX and Process Automation

Statistical data and metadata exchanges employ many different automated processes, but some are of more general interest than others. There are some common information technologies that are nearly ubiquitous within information systems today. SDMX aims to provide standards that are most useful for these automated processes and technologies.

Briefly, these can be described as:

1. *Batch Exchange of Data and Metadata:* The transmission of whole or partial databases between counterparties, including incremental updating.

2. *Provision of Data and Metadata on the Internet:* Internet technology - including its use in private or semi-private TCP/IP networks - is extremely common. This technology includes XML and web services as primary mechanisms for automating data and metadata provision, as well as the more traditional static HTML and database-driven publishing.

3. *Generic Processes:* While many applications and processes are specific to some set of data and metadata, other types of automated services and processes are designed to handle any type of statistical data and metadata whatsoever. This is particularly true in cases where portal sites and data feeds are made available on the Internet.

4. *Presentation and Transformation of Data:* In order to make data and metadata useful to consumers, they must support automated processes that transform them into application-specific processing formats, other standard formats, and presentational formats. Although not strictly an aspect of exchange, this type of automated processing represents a set of requirements that must be supported if the information exchange between counterparties is itself to be supported.

The SDMX standards specified here are designed to support the requirements of all of these automation processes and technologies.

3.3 Statistical Data and Metadata

To avoid confusion about which "data" and "metadata" are the intended content of the SDMX formats specified here, a statement of scope is offered. Statistical "data" are sets of often numeric observations which typically have time associated with them. They are associated with a set of metadata values, representing specific concepts, which act as identifiers and descriptors of the data. These metadata values and concepts can be understood as the named dimensions of a multi-dimensional co-ordinate system, describing what is often called a "cube" of data.

SDMX identifies a standard technique for modelling, expressing, and understanding the structure of this multi-dimensional "cube", allowing automated processing of data from a variety of sources. This approach is widely applicable across types of data and attempts to provide the simplest and most easily comprehensible technique that will support the exchange of this broad set of data and related metadata.

The term "metadata" is very broad indeed. A distinction can be made between “structural” metadata – those concepts used in the description and identification of statistical data and metadata – and “reference” metadata – the larger set of concepts that describe and qualify statistical data sets and processing more generally, and which are often associated not with specific observations or series of data, but with entire collections of data or even the institutions which provide that data.

The SDMX Information Model provides for the structuring not only of data, but also of “reference” metadata. While these reference metadata structures exist independent of the data and its structural metadata, they are often linked. The SDMX Information Model provides for the attachment of reference metadata to any part of the data or structural metadata, as well as for the reporting and exchange of the reference metadata and its structural descriptions. This function of the SDMX standards supports many aspects of data quality initiatives, allowing as it does for the exchange of metadata in its broadest sense, of which quality-related metadata is a major part.

Metadata are associated not only with data, but also with the process of providing and managing the flow of data. The SDMX Information Model provides for a set of metadata concerned with “data provisioning” – metadata which are useful to those who need to understand the content and form of a data provider’s output. Each data provider can describe in standard fashion the content of and dependencies within the data and metadata sets which they produce, and supply information about the scheduling and mechanism by which their data and metadata are provided. This allows for automation of some validation and control functions, as well as supporting management of data reporting.

SDMX also recognizes the importance of classification schemes in organizing and managing the exchange and dissemination of data and metadata. It is possible to express information about classification schemes and domain categories in SDMX, along with their relationships to data and metadata sets, as well as to categorize other objects in the model.

The SDMX standards offer a common model, a choice of syntax and, for XML, a choice of data formats which support the exchange of any type of statistical data meeting the definition above; several optimized formats are specified based on the specific requirements of each implementation, as described below in the SDMX-ML section.

The formal objects in the information model are presented briefly below, but are also discussed in more detail elsewhere in this specification.

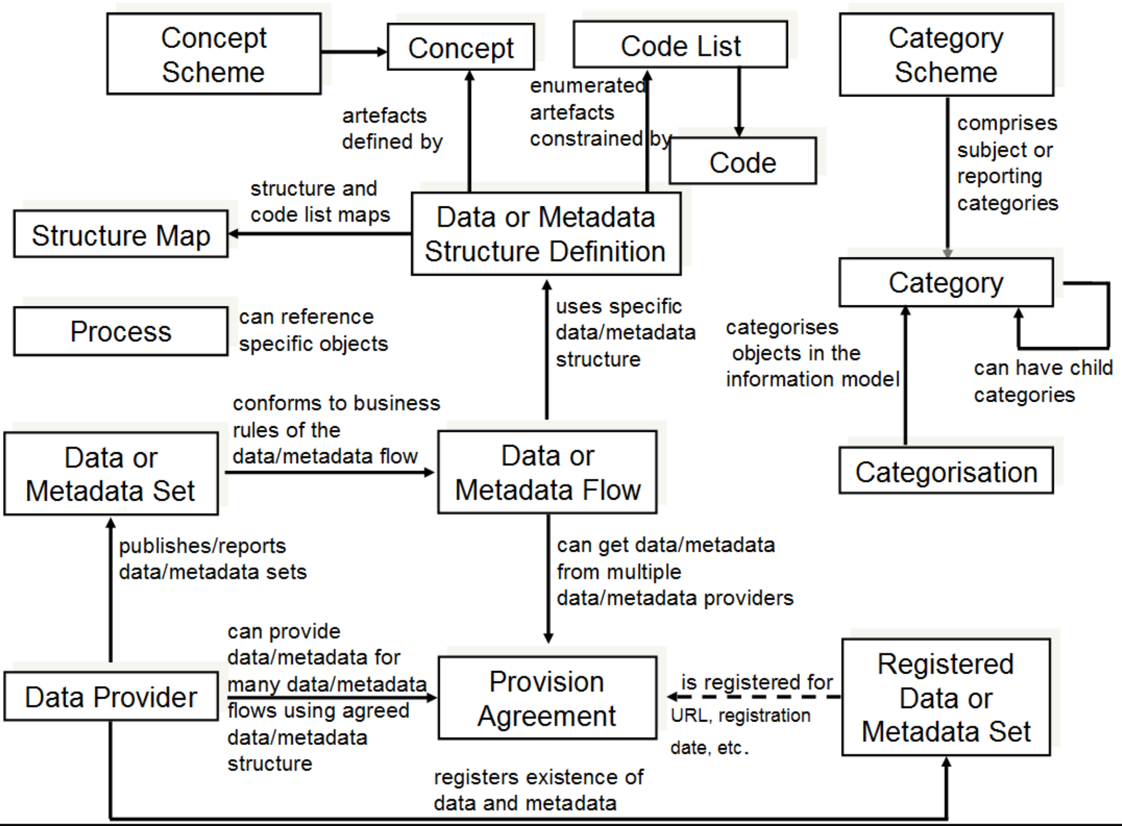

Figure 1: High Level Schematic of Major Artefacts in the SDMX Information Model

3.4 The SDMX View of Statistical Exchange

Version 1.0 of ISO/TS 17369 SDMX covered statistical data sets and the metadata related to the structure of these data sets. This scope was useful in supporting the different models of statistical exchange (bilateral exchange, gateway exchange, and data-sharing) but was not by itself sufficient to support them completely. Versions 2.0 and 2.1 provide a much more complete view of statistical exchange, so that an open data-sharing model can be fully supported, and other models of exchange can be more completely automated. In order to produce technical standards that will support this increased scope, the SDMX Information Model provides a broader set of formal objects which describe the actors, processes, and resources within statistical exchanges.

It is important to understand the set of formal objects not only in a technical sense, but also in terms of what they represent in the real-world exchange of statistical data and metadata.

The first version of SDMX provided for data sets - specific statistical data reported according to a specific structure, for a specific time range - and for data structure definitions - the metadata which describes the structure of statistical data sets. These are important objects in statistical exchanges, and are retained and enhanced in the second version of the standards in a backward-compatible form. A related object in statistical exchanges is the "data flow" - this supports the concept of data reporting or dissemination on an ongoing basis. "Data flows" can be understood as data sets which are not bounded by time. Data structures are owned and maintained by agencies - in a similar fashion, data flows are owned by maintenance agencies.

Versions 2.0 and 2.1 – like version 1.0 – allow for the publication of statistical data (and the related structural metadata) but also provide for the standard, systematic representation of reference metadata. Reference metadata are reported not as an integral part of a data set, but independent of the statistical data. SDMX provides for reference "metadata sets", "metadata structure definitions", and "metadata flows". These objects are very similar to data sets, data structure definitions, and data flows, but they concern reference metadata rather than statistical observations. In the same way that data providers may publish statistical data, they may also publish reference metadata. Metadata structural definitions are maintained by agencies in a fashion similar to the way that agencies maintain data structure definitions, the structural definitions of data sets.

The structural definitions of both data and reference metadata associate specific statistical concepts with their representations, whether textual, coded, etc. In SDMX version 2.0/2.1, these concepts are taken from a "concept scheme" which is maintained by a specific agency. Concept schemes group a set of concepts, provide their definitions and names, and allow for semantic relationships to be expressed, when some concepts are specializations of others. It is possible for a single concept scheme to be used both for data structures - key families - and for reference metadata structures.

Inherent in any statistical exchange – and in many dissemination activities - is a concept of "service level agreement", even if this is not formalized or made explicit. SDMX incorporates this idea in objects termed "provision agreements". Data providers may provide data to many different data flows. Data flows may incorporate data coming from more than one data provider. Provision agreements are the objects which tell you which data providers are supplying what data to which data flows. The same is true for metadata flows.

Provision agreements allow for a variety of information to be made available: the schedule by which statistical data or metadata is reported or published, the specific topics about which data or metadata is reported within the theoretically possible set of data (as described by a data structure definition or reference metadata structure definition), and the time period covered by the statistical data and metadata. This set of information is termed "constraint" in the SDMX Information Model.

A brief summary of the objects described in the information model includes:

- ***Data Set:*** Data is organized into discrete sets, which include particular observations for a specific period of time. A data set can be understood as a collection of similar data, sharing a structure, which covers a fixed period of time.

- ***Data Structure Definition (DSD, also known as Key Family in Version 2.0):*** Each data set has a set of structural metadata. These descriptions are referred to in SDMX as Data Structure Definitions, which include information about how concepts are associated with the measures, dimensions, and attributes of a data "cube," along with information about the representation of data and related identifying and descriptive (structural) metadata. In Version 2.1, the term "Key Family" is replaced by "Data Structure Definition" (DSD) both in XML Schemas and the Information Model.
- ***Code list:*** Code lists enumerate a set of values to be used in the representation of dimensions, attributes, and other structural parts of SDMX. They can be supplemented by other structural metadata which indicates how codes are organized into hierarchies.
- ***Organisation Scheme:*** Organisations and organisation structure can be defined in an Organisation Scheme. Specific Organisation Schemes exist for Maintenance Agency, Data Provider, Data Consumer, and Organisation Unit.
- ***Category Scheme and Categorisation:*** Category schemes are made up of a hierarchy of categories, which in SDMX may include any type of useful classification for the organization of data and metadata. A Categorisation links a category to an identifiable object. In this way sets of objects can be categorised. A statistical subject-matter domain scheme is implemented in SDMX as a Category Scheme.
- ***Concept Scheme:*** A concept scheme is a maintained list of concepts that are used in data structure definitions and metadata structure definitions. There can be many such concept schemes. A "core" representation of the concept can be specified (e.g. a core code list, or other representation such as "date"). Note that this core representation can be overridden in the data structure definition or metadata structure definition that uses the concept. Indeed, organisations wishing to remain with version 1.0 key family schema specifications will continue to declare the representation in the key family definition.
- ***Metadata Set:*** A reference metadata set is a set of information pertaining to an object within the formal SDMX view of statistical exchange: they may describe the maintainers of data or structural definitions; they may describe the schedule on which data is released; they may describe the flow of a single type of data over time; they may describe the quality of data, etc. In SDMX, the creators of reference metadata may take whatever concepts they are concerned with, or obliged to report, and provide a reference metadata set containing that information.
- ***Metadata Structure Definition:*** A reference metadata set also has a set of structural metadata which describes how it is organized. This metadata set identifies what reference metadata concepts are being reported, how these concepts relate to each other (typically as hierarchies), what their presentational structure is, how they may be represented (as free text, as coded values, etc.), and with which formal SDMX object types they are associated.
- ***Dataflow Definition:*** In SDMX, data sets are reported or disseminated according to a data flow definition. The data flow definition identifies the data structure definition and may be associated with one or more subject matter domains via a Categorisation (this facilitates the search for data according to

organised category schemes). Constraints, in terms of reporting periodicity or sub set of possible keys that are allowed in a data set, may be attached to the data flow definition.

- ***Metadataflow Definition:*** A metadata flow definition is very similar to a data flow definition, but describes, categorises, and constrains metadata sets.
- ***Data Provider:*** An organization which produces data or reference metadata is termed a data provider.
- ***Provision Agreement:*** The set of information which describes the way in which data sets and metadata sets are provided by a data provider. A provision agreement can be constrained in much the same way as a data or metadata flow definition. Thus, a data provider can express the fact that it provides a particular data flow covering a specific set of countries and topics, Importantly, the actual source of registered data or metadata is attached to the provision agreement (in terms of a URL). The term "agreement" is used because this information can be understood as the basis of a "service-level agreement". In SDMX, however, this is informational metadata to support the technical systems, as opposed to any sort of contractual information (which is outside the scope of a technical specification).
- ***Constraint:*** Constraints describe a subset of a data source or metadata source, and may also provide information about scheduled releases of data. They are associated with data providers, provision agreements, data flows, metadataflows, data structure definitions and metadata structure definitions.
- ***Structure Set:*** Structure sets provide a mechanism for grouping structural metadata together to form a complete description of the relationships between specific, related sets of data and metadata. They can be used to map dimensions and attributes to one another, to map concepts, to map code lists, and to map category schemes. They can be used to describe "cubes" of data, even when the data within the cube does not share a single dimensionality.
- ***Reporting Taxonomy:*** A reporting taxonomy allows an organisation to link (possibly in a hierarchical way) a number of cube or data flow definitions which together form a complete "report" of data or metadata. This supports primary reporting which often comprises multiple cubes of heterogeneous data, but may also support other collection and reporting functions. It also supports the specification of publications such as a yearbook, in terms of the data or metadata contained in the publication.
- ***Process:*** The process class provides a way to model statistical processes as a set of interconnected *process steps.* Although not central to the exchange and dissemination of statistical data and metadata, having a shared description of processing allows for the interoperable exchange and dissemination of reference metadata sets which describe processes-related concepts.
- ***Hierarchical Code List:*** This supports the specification of code hierarchies. The codes themselves are referenced from the code lists in which they are maintained. The Hierarchical Code List thus specifies the organisation of the codes in one or more hierarchies, but does not define the codes themselves.

Notes on Data Structuring

A "cube" is a rich, multi-dimensional construct, which can be viewed along any of its axes (or "dimensions"). Whilst the full structure of cube data can be described in

SDMX, the actual “data” specification of SDMX takes a slightly narrower view of these requirements in its version 2.0/2.1 specifications for the purposes of formatting the data for transmission. The view of data in many SDMX formats is primarily as time series – that is, as a set of observations which are organized around the time dimension, so that each observation occurs progressively through time.

There are, however, many types of statistical data which are not typically organized for exchange as time series where data are organized around some other, non-time dimension of the cube – what is often called “cross-sectional” data. SDMX supports a unified format that represents in the data set an organisation of the data along any single dimension. In this context, time series is a particular case of the unified format.

Another type of structure commonly found in statistical “cubes” of data is the hierarchical classification, used to describe the points along any of its dimensions (or axes). In the 1.0 version, SDMX standards did not provide full support for this functionality. The introduction of these hierarchical classifications is present in the current version of the standard.

Further, there is support for the expression of functional dependencies between the various dimensions of a cube, giving support for better processing of “sparse cubes”. This is an aspect of “constraints”, which allow for the framing of a cube region, or for the provision of a set of valid keys within the total set of keys described by the data structure definition.

Notes on Reference Metadata Structuring

Metadata structures are based on the idea that concepts can be organised into semantic and presentational hierarchies, and that these hierarchies can form the basis for the structuring of XML reporting formats. There are three message types in SDMX-ML which serve this purpose: the Structure message (providing the metadata structure definition), the Generic Metadata message (providing a single format for any metadata structure definition), and the Structure-specific Metadata message (providing a metadata structure definition-specific format). Typically, this mechanism is suited to supporting reference metadata reporting and dissemination.

The Metadata Structure Definition takes *any* concept from concept schemes, and describes how they can be formed into a reporting or dissemination structure as metadata attributes – either as a flat list, or as a hierarchy. The metadata attributes are assigned representations (coded, textual, etc.) and the number of occurrences. The “target” of the metadata – that is, the class of process, information, organisation, exchange, etc. – which is the subject of the metadata is described. Because the SDMX Information Model gives a formalization of statistical exchange and dissemination, the model can be used as a typology of the different actors and resources within statistical activities. Thus, the “targets” (subjects) of reference metadata sets and metadata flows can be described as corresponding to some standard class by reference to this model.

As with data structures, the generic format for metadata sets provides a known document structure, whilst the structure specific format is derived specifically from a metadata structure definition and can perform a higher degree of schema validation.

3.5 SDMX Registry Services

In order to provide visibility into the large amount of data and metadata which exists within the SDMX model of statistical exchange, it is felt that an architecture based on a set of registry services is potentially useful. A “registry” – as understood in web-services terminology – is an application which maintains and stores metadata for querying, and which can be used by any other application in the network with sufficient access privileges (though note that the mechanism of access control is outside of the scope of the SDMX standard). It can be understood as the index of a distributed database or metadata repository which is made up of all the data provider’s data sets and reference metadata sets within a statistical community, located across the Internet or similar network.

Note that the SDMX registry services are not concerned with the storage of data or reference metadata. The assumption is that data and reference metadata lives on the sites of its data providers. The SDMX registry services concern themselves with providing visibility of the data and reference metadata, and information needed to access the data and reference metadata. Thus, a registered data set will have its URL available in the registry, but not the data itself. An application which wishes to access that data would query the registry, perhaps by drilling down via a Category Scheme and Dataflow, for the URL of a registered data source, and then retrieve the data directly from the data provider (using an SDMX-ML query message or other mechanism).

SDMX does not require a particular technology implementation of the registry – instead, it specifies the standard interfaces which may be supported by a registry. Thus, users may implement an SDMX-conformant registry in any fashion they choose, so long as the interfaces are supported as specified here. These interfaces are expressed as XML documents, and form a new part of the SDMX-ML language.

The registry services discussed here can be briefly summarized:

- ***Maintenance of Structural Metadata*:** This registry service allows users with maintenance agency access privileges to submit and modify structural metadata. In this aspect the registry is acting as a structural metadata repository. However, it is permissible in an SDMX structure to submit just the “stub” of the structural object, such as a code list, and for this stub to reference the actual location from where the metadata can be retrieved, either from a file or a structural metadata resource, such as another registry.
- ***Registration of Data and Metadata Sources:*** This registry service allows users with maintenance agency access privileges to inform the registry of the existence and location (for retrieval) of data sets and reference metadata sets. The registry stores metadata about these objects, and links it to the structural metadata that give sufficient structural information for an application to process it, or for an application to discover its existence. Objects in the registry are organized and categorized according to one or more category schemes.
- ***Querying:*** The registry services have interfaces for querying the metadata contained in a registry, so that applications and users can discover the existence of data sets and reference metadata sets, structural metadata, the providers/agencies associated with those objects, and the provider

agreements which describe how the data and metadata are made available, and how they are categorized.

- ***Subscription/Notification:*** It is possible to “subscribe” to specific objects in a registry, so that a notification will be sent to all subscribers whenever the registry objects are updated.

3.6 Web services

Web services allow computer applications to exchange data directly over the Internet, essentially allowing modular or distributed computing in a more flexible fashion than ever before. In order to allow web services to function, however, many standards are required: for requesting and supplying data; for expressing the enveloping data which is used to package exchanged data; for describing web services to one another, to allow for easy integration into applications that use other web services as data resources.

SDMX provides guidelines for using these standards in a fashion which will promote interoperability among SDMX web services, and allow for the creation of generic client applications which will be able to communicate meaningfully with any SDMX web service which implements these guidelines.

More specifically, the SDMX web services guidelines offer:

- A normative interface (WSDL) for SOAP-based web services: The 2.0 Web-Services Guidelines contained a set of web-services functions, but these have been found through implementation to be insufficient for the types of SDMX-based web services now being developed. Furthermore, the operations and their payload have now become normative (WSDL).

- A normative interface (WADL) for RESTful web services: The RESTful API focuses on simplicity. The aim is not to replicate the full semantic richness of the SDMX-ML Query message but to make it simple to perform a limited set of standard queries. Also, in contrast to other parts of the SDMX specification, the RESTful API focuses solely on data retrieval (via HTTP GET).

A normative list of common error codes: When web services are used, it is necessary to have error codes which can help to explain the situation when problems are encountered. Prior to version 2.1 of the SDMX standard, there was no set of agreed error codes for use with SDMX web services. Version 2.1 of the SDMX standard fills that gap.

4 The SDMX Information Model

SDMX provides a way of modelling statistical data, and defines the set of metadata constructs used for this purpose. Because SDMX specifies formats in two syntaxes for expressing data and structural metadata, the model is used as a mechanism for guaranteeing that transformation between the different formats are lossless. All of the formats are syntax-bound expressions of the common information model. SDMX version 1.0 has based itself on GESMES/TS as an input to the model and formats, both to build on the proven success of this model for time series data exchange, and to ensure backward compatibility with existing GESMES/TS-based systems. Version 2.0/2.1 expands upon the version 1.0 basis to provide a more comprehensive model.

SDMX recognizes that statistical data is structured; in SDMX this structure is termed a Data Structure Definition. “Data sets” are made up of one or more lower-level “groups”, based on their degrees of similarity. Each group is in turn comprised of one or more “series” of data. Each series or section has a “key” - values for each of a cluster of concepts, also called "dimensions" - which identifies it, and one or more “observations”, which typically combine the time of the observation, and the value of the observation (e.g., measurement). Additionally, metadata may be attached at any level of this structure as descriptive “attributes”. Code lists (enumerations) and other patterns for representation of data and metadata are also modelled.

There is some similarity between “cube” structures commonly used to process statistical data, and the Data Structure Definition idea in the SDMX Information Model. It is important to note that the data as structured according to the SDMX Information Model is optimized for exchange, potentially with partners who may have no ability to process a “cube” of data coming from complex statistical systems. SDMX time series can be understood as “slices” of the cube. Such a slice is identified by its key. A "series" key consists of the values for all dimensions specified by the key family except time. It is certainly possible to reconstruct and describe data cubes from SDMX-structured data, and to exchange such databases according to the proposed standards. In version 2.0, it becomes possible to more fully describe the structure of cubes, with hierarchical code lists, constraints, and relationships between data structure definitions.

In version 2.0/2.1, the SDMX standards also provide a view of reference metadata: a mechanism for referencing the meaningful “objects” within the SDMX view of statistical exchange processes (data providers, structures, provisioning agreements, dataflows, metadata flows, etc.) to which metadata is attached; a mechanism for describing a set of meaningful concepts, of organizing them into a presentational structure, and of indicating how their values are represented. This is based on a simple, hierarchical view of reference metadata which is common to many metadata systems and classification/categorization schemes. SDMX provides a model (and XML formats) for both describing reference metadata structures, and of reporting reference metadata according to those structures.

Version 2.0/2.1 also introduces support for metadata related to the process aspects of statistical exchange. A step-by-step process can be modelled; information about who is providing data and reference metadata and how they are providing it can be expressed; and the technical aspects of service-level agreements (and similar types of provisioning agreements) can be represented.

The SDMX Information Model formally describes all of the objects listed above, so as to present a standard view of the statistical exchange process.

The SDMX Information Model is presented using UML, and is also described in prose. While the information model is not normative, it is a valuable tool for understanding and using the normative format specifications.

5 SDMX-EDI

The SDMX-EDI format is drawn from the GESMES/TS version 3.0 implementation guide, as published as a standard of the SDMX initiative.

1. *Statistical Definitions:* An expression of the structural metadata covered by the SDMX information model in a UN/EDIFACT format.

2. *Statistical Data:* Optimized for the batch exchange of large amounts of time series data between counterparties, it allows for extremely compact expression of large whole or partial data sets. Non time series data, such as cross-sectional, can be supported if represented as repackaged time series, but there is no direct support for cross-sectional data in this format.

3. *Data Set List:* a list of data sets and their structural metadata.

The SDMX Information Model provides the constructs which are found in the EDIFACT syntax used for SDMX-EDI, and those found in the XML syntax of SDMX-ML. Since both syntactic implementations reflect the same logical constructs, SDMX-EDI data and structural metadata messages can be transformed into corresponding SDMX-ML formats, and vice-versa. Thus, these standards provide for interoperability between the UN/EDIFACT-based and XML-based systems processing and exchanging statistical data and metadata.

6 SDMX-ML

While the SDMX-EDI format is primarily designed to support batch exchange, SDMX-ML supports a wider range of requirements. XML formats are used for many different types of automated processing, and thus must support more varied processing scenarios. That is why there are several types of messages available as SDMX-ML formats. Each is suited to support a specific set of processing requirements.

1. *Structure Definition:* All SDMX-ML message types share a common XML expression of the metadata needed to understand and process a data set or metadata set, and additional metadata about category schemes and organisations is included. Also, the structural aspects of data and metadata provision – dataflows and metadataflows – are described using this format.

2. *Generic Data:* All statistical data expressible in SDMX-ML can be marked up according to this data format, in agreement with the contents of a Structure Definition message. It is designed for any scenario where applications receiving the data need to process it according to a single format. Such applications may need independent access to the data set's structure before they process it. Data marked up in this format are not particularly compact, but they make easily available all aspects of the data set. This format does not provide strict validation between the data set and its structural definition using a generic XML parser. It supports the transmission of partial data sets (incremental updates) as well as whole data sets. It supports both the time-series and the cross-sectional use cases.

3. *Structure-specific Data:* This format is specific to the Data Structure Definition of the data set (in other terms, it is DSD-specific) and is created by following mappings between the metadata constructs defined in the Structure Definition message and the technical specification of the format. It supports the exchange of large data sets in XML format (typically the size of the data set is 50% of the same data expressed as Generic Data), provides strict validation

of conformance with the DSD using a generic XML parser, and supports the transmission of partial data sets (incremental updates) as well as whole data sets. The Structure-specific Data format specified in SDMX 2.1 supports both the time-series and the cross-sectional use cases which were covered by two distinct formats in SDMX 2.0.

Many XML tools and technologies have expectations about the functions performed by an XML schema, one of which is a very direct relationship between the XML constructs described in the XML schema and the tagged data in the XML instance. Strong data typing is also considered normal, supporting full validation of the tagged data. These message types are designed to support validation and other expected XML schema functions.

4. *Generic Metadata:* All reference metadata expressible in SDMX-ML format can be marked up according to this schema. It performs only a minimum of validation, and is somewhat verbose, but it does support the creation of generic software tools and services for processing reference metadata.

5. *Structure-specific Metadata:* For each metadata structure definition, an XML schema specific to that structure can be created, to perform validation on sets of reported metadata. This structure is less verbose than the Generic Metadata format, and, because the XML mark-up relates directly to the reported concepts, it is appropriate for applications that are designed to process a specific type of metadata report. It is analogous to the Structure-specific Data format for data in its approach to the use of XML.

6. *Query:* Data and metadata are often published in databases which are available on the web. Thus, it is necessary to have a standard query document which allows the databases to be queried, and return an SDMX-ML data, reference metadata, or structure message. The Query document is an implementation of the SDMX Information Model for use in web services and database-driven applications, allowing for a standard request to be sent to data providers using these technologies.

7. *Registry:* All of the possible interactions with the SDMX registry services are supported using SDMX-ML interfaces. All but one of these documents are based on a synchronous exchange of documents – a “request” message answered by a “response” message. There are two basic types of request – a “Submit”, which writes metadata to the registry services, and a “Query”, which is used to discover that metadata. Registry interactions provide formats for all types of provisioning metadata, as well as for subscription/notification, structural metadata, and data and metadata registration. The exception is the (Registry) notification message which is asynchronous.

Because all of the SDMX-ML formats are implementations of the same information model, and all the data and metadata messages are derivable from the Structure message which describes a data set or metadata set, it is possible to have standard mappings between each of the similar formats. These mappings can be implemented in generic transformation tools, useful to all SDMX-ML users, and not specific to a particular data set’s key family or metadata set’s structure definition (even though some of the formats they deal with may be). Part of the SDMX-ML package is the set of mappings between the structure-specific data and metadata formats and the Structure Definition format from which all are derivable.

7 Conformance

This section will contain a normative statement of what applications must do to be considered conformant with the SDMX version 2.1 specifications. This will address both the application functionality that must be supported, and the contents of an Implementer's Conformance Statement regarding SDMX conformance.

8 Dependencies on SDMX content-oriented guidelines

The technical standards proposed here are designed so that they can be used in conjunction with other SDMX guidelines which are more closely tied to the content and semantics of statistical data exchange. The SDMX Information Model works equally well with any statistical concept, but to encourage interoperability, it is also necessary to standardize and harmonize the use of specific concepts and terminology. To achieve this goal, SDMX creates and maintains guidelines for cross-domain concepts, terminology, and structural definitions. There are three major parts to this effort.

8.1 Cross-Domain Concepts

The SDMX Cross-Domain Concepts is a content guideline concerning concepts which are used across statistical domains. This list is expected to grow and to be subject to revision as SDMX is used in a growing number of domains. The use of the SDMX Cross-Domain Concepts, where appropriate, provides a framework to further promote interoperability among organisations using the technical standards presented here. The harmonization of statistical concepts includes not only the definitions of the concepts, and their names, but also, where appropriate, their representation with standard code lists, and the role they play within data structure definitions and metadata structure definitions.

The intent of this guideline is two-fold: to provide a core set of concepts which can be used to structure statistical data and metadata, to promote interoperability between systems ("structural metadata", as described above); and to promote the exchange of metadata more widely, with a set of harmonized concept names and definitions for other types of metadata ("reference metadata", as defined above.)

8.2 Metadata Common Vocabulary

The Metadata Common Vocabulary is an SDMX guideline which provides definition of terms to be used for the comparison and mapping of terminology found in data structure definitions and in other aspects of statistical metadata management. Essentially, it provides ISO-compliant definitions for a wide range of statistical terms, which may be used directly, or against which other terminology systems may be mapped. This set of terms is inclusive of the terminology used within the SDMX Technical Standards.

The MCV provides definitions for terms on which the SDMX Cross-Domain Metadata Concepts work is built.

8.3 Statistical Subject-Matter Domains

The Statistical Subject-Matter Domains is a listing of the breadth of statistical information for the purposes of organizing widespread statistical exchange and categorization. It acts as a standard scheme against which the categorization schemes of various counterparties can be mapped, to facilitate interoperable data and metadata exchange. It serves another useful purpose, however, which is to allow an organization of corresponding "domain groups", each of which could define standard data structure definitions, concepts, etc. within their domains. Such groups already exist within the international community. SDMX would use the Statistical Subject-Matter Domains list to facilitate the efforts of these groups to develop the kinds of content standards which could support the interoperation of SDMX-conformant technical systems within and across statistical domains. The organisation of the content of such schemes is supported in SDMX as a Category Scheme.

SDMX Statistical Subject-Matter Domains will be listed and maintained by the SDMX Initiative and will be subject to adjustment.

9 Looking Forward

The SDMX initiative sees this set of data and metadata formats and registry services interfaces standards as useful in creating more efficient and open systems for statistical exchange. It is anticipated that SDMX will refine these standards further as they are implemented, so as to build on the interoperability enabled by having a set of standard formats and exchanges based on a common information model.

The review process for version 2.0 and 2.1 has suggested that future work should take advantage of a wider participation of the SDMX user community (statistical offices, central banks and other national and international organisations dealing with statistics) in further enhancing the Technical Standards and improving its use.