**The expanding role of metadata supporting a modern statistical production process**

Essi Kaukonen, Statistics Finland, essi.kaukonen@stat.fi

Mikko Saloila, Statistics Finland, mikko.saloila@stat.fi

**Abstract**

*The business environment of National Statistical Offices (NSOs) is changing. We are facing the challenge of adapting quicker to new information needs. We are also required to produce reliable information spending less time on the production process. These drivers lead to new demands for producing data resulting in an expanded role for metadata.*

*Today, metadata needs to drive processes, provide traceability, as well as support extensive usability of data inside and outside NSOs. In order to drive processes, metadata needs to be not only machine-readable but also uniform and designed to meet the refined, often data structure related, requirements. To provide traceability it is required to manage identified data objects and associated lineages in a coherent way. In order to track complex data transformations, well designed services need to be created. Usability support is the traditional role usually discussed when talking about statistical metadata. The new requirements involving this role are related to the level of granularity and ability to reuse core metadata across subject fields and production phases.*

*There are two main international standards in this area: The Generic Statistical Information Model (GSIM) and the Generic Statistical Business Process Model (GSBPM). These frameworks governed by the United Nations Economic Commission for Europe (UNECE) are crucial to enable development around modern statistical data management to be applicable to different NSOs. Statistics Finland currently has ongoing projects working with metadata driven processes. We are also renewing our metadata system to be GSIM-based and to support the new roles of metadata. Still, a lot of open questions remain in order to meet all the new requirements. These issues would benefit greatly from collaboration within the statistical community.*

**Keywords:** Modernising of statistics, GSBPM, GSIM, Metadata-driven, Data management

**1. Changes in the business environment of National Statistical Offices**

The business environment of National Statistical Offices has been changing in many ways. The digital age, as well as new and renewing data sources, affect the production of official statistics. A broader range of users are potentially using official statistics and demanding a broader range of statistical products. Our customers are asking for closer to real time information concerning very specific issues. On the other hand, the need for internationally comparable global information has also increased. At the same time with increasing demands and new environments, the NSOs are facing a reducing level of resources. (*Dalton et al. 2015.)*

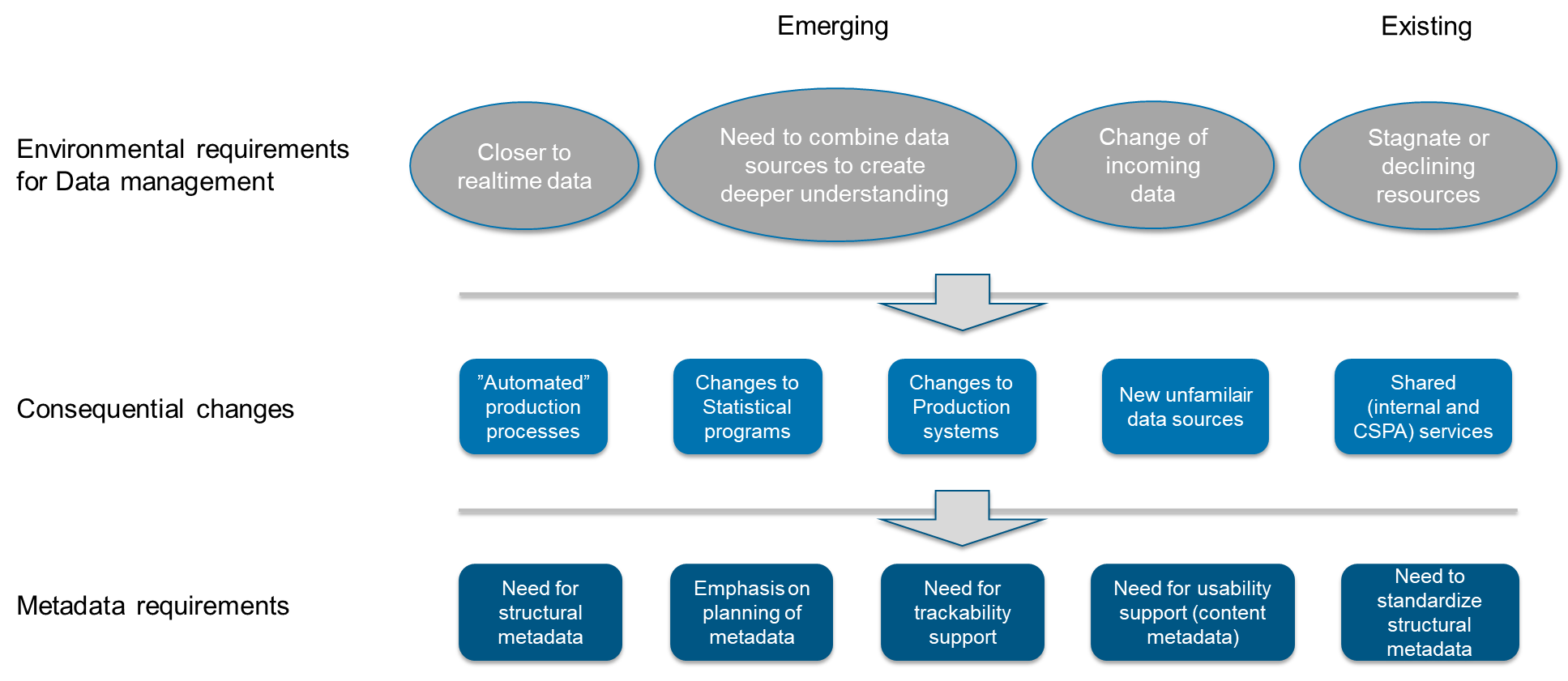
This environmental change also applies to the Nordic countries and Statistics Finland. To meet this change, a new strategy has been formed for Statistics Finland.

**Figure 1. New strategy of Statistics Finland: Chosen factors of change**

  
Source: Statistics Finland strategy, 2019

The new strategy has a significant impact on data management issues, as well as on interoperability within the public sector. In order to implement our strategy successfully, we must consider what the effects on our current Data management policies and processes are. In Figure 2, we identify some of the key environmental requirements for Data management, the consequential changes to statistical production and the following metadata requirements.

**Figure 2. Requirements for data management and the consequential changes**



One concrete use case for new and changing administrative data sources is the Incomes Register from the Tax Administration or Right to study and achievement -register from the Finnish National Agency for Education. These registers have traditionally provided data monthly or yearly. In the near future, this data is provided more frequently, even daily, and usually consists of events. These kind of data sources could be an answer to the need to get closer to real time data but, at the same time, cause substantial changes to statistical programs and production systems, as well as creating new metadata requirements.

**2. Metadata management eras at Statistics Finland**

*2.1. Metadata definitions in NSO context*

Metadata is an awfully broad term. Generally, it is described as data about data. A more specific definition of metadata can be found, for example, in the Finnish thesaurus and ontology service: information that describes the context, contents or structure of a piece of data and also helps guide and document its utilisation and management (source: Information Terms glossary).

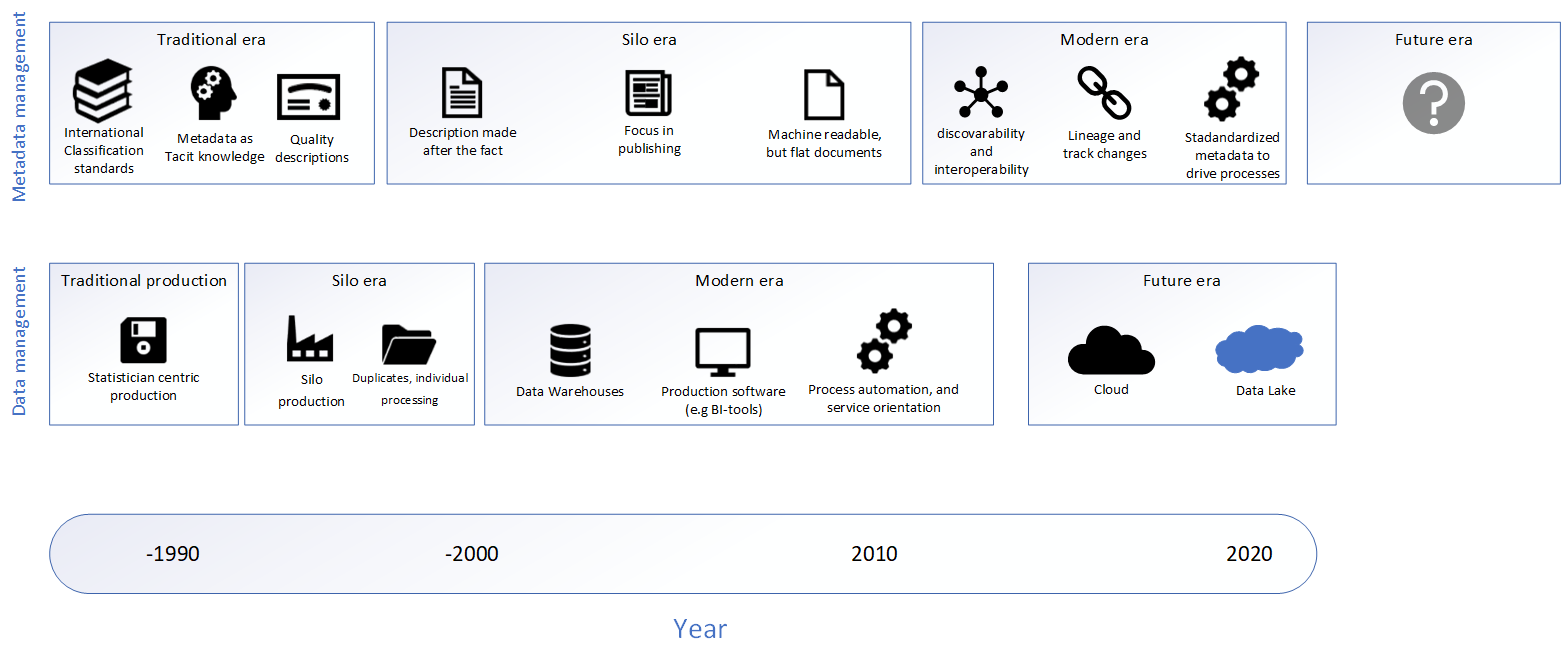
In the statistical domain, perhaps the most used definition of metadata is found in the SDMX glossary. SDMX divides metadata into two concepts, Reference metadata and Structural metadata. Reference metadata is defined as “Metadata describing the contents and the quality of the statistical data”. Structural metadata is described as “Metadata that identify and describe data and reference metadata” (SDMX Glossary, 2018a). SDMX was originally developed for data exchange and, therefore, metadata related to process, which will be referenced to later in this paper, is not included in this definition.

To make things somewhat easier in this paper, we restrict our examination to information objects needed in statistical production. These objects are defined in the GSIM information model (United Nations Economic Commission, 2019a). We use this framework to stress the importance and overall nature of metadata for statistical production.

*2.2 A brief history of Metadata management at Statistics Finland*

Metadata is not an isolated issue. Metadata management is, in fact, tightly tied to Data management, which takes care of the actual data needed for official statistics and other information products created by NSOs. In an ideal world these two, Data management and Metadata management, would be led and developed in collaboration. However, the reality at Statistics Finland has been far from joined Data and Metadata management, especially during past decades. In figure 2, the Data management eras and Metadata management eras are pictured parallel; Metadata management is clearly lagging behind.

**Figure 3. Metadata and data management eras at Statistics Finland**



When we have a closer look at the” Traditional era” of Metadata management in the NSO of Finland, we clearly see that the focus lied on taking care of international classification standards created in international cooperation. The use of these standards was promoted especially inside our NSO, but for the stakeholders outside our organization these were, at least in a machine-readable format, subject to a charge. During this period, statistical production was pretty much statistician centric, meaning that content metadata used inside our NSO was mainly accessible to the statisticians themselves working for a certain Statistical Program (United Nations Economic Commission, 2019a). Metadata could even be in a non-written format in those days. This content metadata was not published to our customers, whereas quality information was disseminated first asproduct descriptions until 2007, and then continued as Quality descriptions when an XML-based publication system was implemented at Statistics Finland. These Quality descriptions had more content and followed the European Statistics Code of Practice (Statistics Finland, 2014)

In the 1990s, we moved from the Traditional era of Metadata management to another era which could be called the Silo era. Statistical production as such started to move forward from the Silo era to a Modern era, but this change only concerned metadata years later. Typical for this Silo era of Metadata management was that the descriptions were made afterwards. In practice, metadata was considered as a dissemination task, as dissemination was the first unified production phase in our NSO. Machine-readable format for metadata was already reality, but the metadata documents were “flat” and mainly not connected to each other.

The development of dissemination technologies during the Silo era eventually led to publishing statistical products on the Internet. This also influenced the development of Statistics Finland’s information model called CoSSI (Common Structure of Statistical Information). The model was created at the beginning of the 21st century. The thinking behind this and the idea of a common information model for an NSO was very advanced for the time. The main idea was to have a uniform and uniting structural definition of statistical information, that could be used in both production and dissemination of statistics (Rouhujärvi and Lehtinen, 2003). The model was later implemented in the publishing system and statistical metadata system at Statistics Finland. These systems mainly handled information in the dissemination phase of statistical production. However, a lot of content metadata concerning the other phases of production apart from dissemination still existed inside different production systems only available to the experts using these systems. The Statistical Program and Product information was to some extent centrally governed, though not fully as a common asset yet.

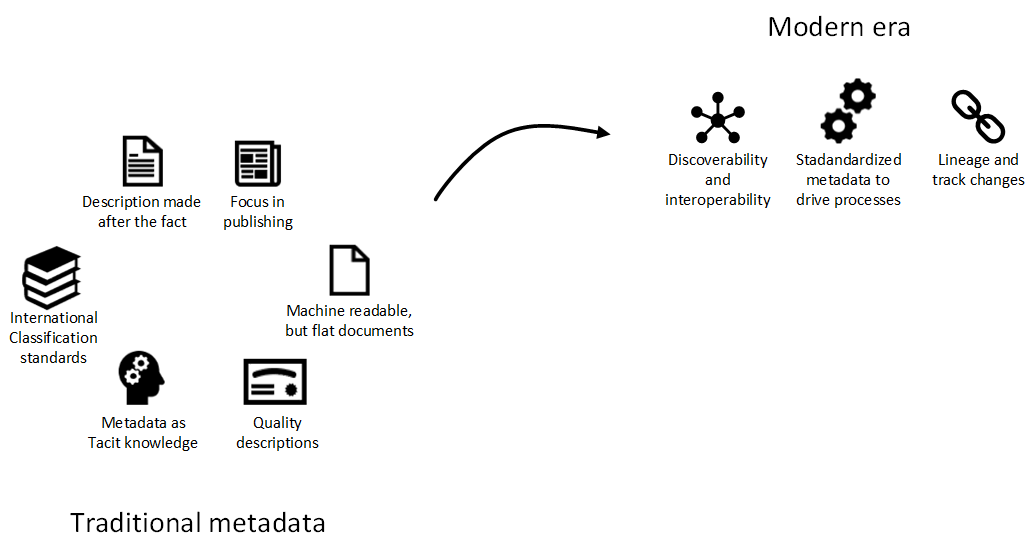
We could say that the Modern era of Metadata management emerged from around 2014 onwards. During this era, several new demanding requirements for metadata have emerged. A relatively new need emerged from the Data management and process management side. The production processes needed to be automated and, hence, metadata was needed to guide these processes. Secondly, the data needed to be shared across different processes and services to enable, for instance, quick response to new customer needs. Furthermore, the metadata used should be standardized at least within an organisation. Finally, it was more and more evident that metadata was also needed to track changes that have been made to the data, as well as information flow.

All in all, metadata and data have developed in different rhythms partly separately from each other. Data is the core asset for an NSO and, consequently, it is in focus when modernising statistical production. Therefore, metadata development is lagging behind and facing the challenge of needing to serve the statistical production of today or even tomorrow with the metadata assets of yesterday*.* Even though the production has largely moved on from silos, Metadata management is, in fact, still largely attached to that era.

**3. Catching up with the modern era in metadata management**

What should be done with metadata in the Modern era? It is not enough to take care of the traditional flat descriptions created for dissemination. Instead, the following characteristics of metadata are essential in order to meet the requirements of the modern era: machine readability, reusability and identifiability. Furthermore, metadata needs to be planned carefully so that it can handle all the functions that are potentially achievable. Metadata can support the usability of data by enhancing discoverability and interoperability. It can also be used to drive processes and track changes.

**Figure 4. Moving from the Traditional era to the Modern era**



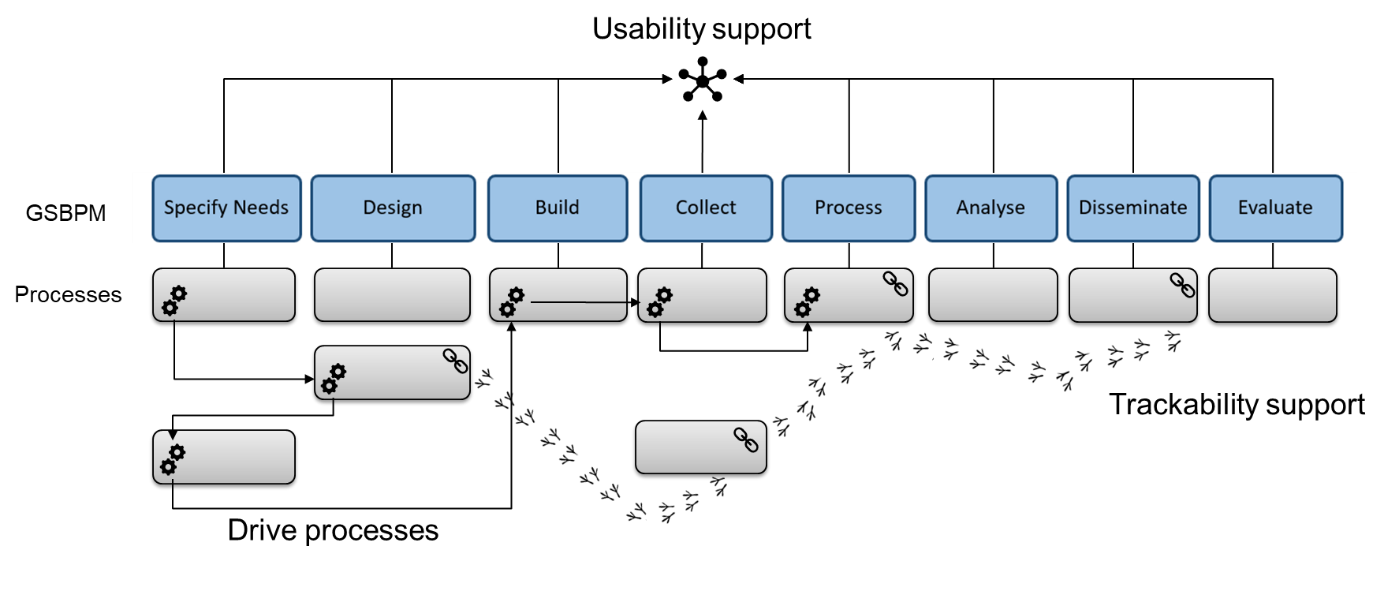
We claim, in fact, that three new roles emerge for metadata in the Modern era. Firstly, in modern statistical production we need to share common data resources across processes and services in order to produce new versatile products and create e.g. new compiled information for our customers faster than before. Improving usability support with metadata is essential here. Metadata is required to support the discoverability and interoperability of the data inside NSOs to better enable services to our customers. To be able to compare the vast amount of data we have today, standardization of the core metadata objects is crucial.

UNECE Metadata Framework recognised, already in 2009, that an effective Metadata System allows organisations to be flexible and responsive to rapidly evolving requirements for statistical information. Well-maintained metadata allows the organisation to operate in a more transparent and quality assured manner. At the same time, organisation-wide metadata supports more effective change management processes, reducing risks to business continuity and the barriers to business process improvement. (United Nations Economic Commission, 2009.)

The focus of the metadata development in the silo era of Metadata management was to make data usable for the users. What is new in this role in today’s challenging environment, is that adaptive data production would need the metadata system to support discoverability and interoperability of the data warehouses used in modern statistical production.

While the basic idea is to support customers inside NSOs, we should not forget the world outside our organisation, the data users and data producers. In fact, today special focus in Finland lies on building information ecosystems and information interoperability. A governmental program, that began in 2016, is aiming to make information interoperable across the public sector. The work is now being continued at the Population Register Centre (Population Register Centre, 2019). Our task as an NSO is to ensure that we succeed in connecting our data with the data produced by other governmental bodies utilising the interoperability solutions and, in practice, linking our metadata to these solutions via APIs and identifiers.

**Figure 5. Metadata roles in modern statistical production**



The second role is related to an increasing need for automated processes. A lot of the questions related to process have already long been solved inside separate production systems. The new thing is to manage data coherently, which requires unified solutions.

Metadata is needed to drive the automated processes to make them truly intelligent. To enable this, structural and granular metadata is required: metadata which, like a mirror, describes even the finest data structures ingested in the modern processes. For process centric experts, this can be the moment when they start, perhaps for the first time, to understand the full potential of metadata. However, this kind of metadata needs to be planned carefully and in a unified way, otherwise the processes will not fit together. For the data coming in and out of these processes the Generic Statistical Information Model (United Nations Economic Commission, 2019a) offers a profound reference point at the conceptual level offering a unified set of recognised concepts. However, in practice, there are several implementation possibilities for logical and physical level data models.

The third role is related to the need to track changes and the increasing flow of data. Lineage is information on the changes that have occurred to the data over its life-cycle (United Nations Economic Commission 2019c). Traceability and lineage are important today because Data management is evolving towards ever bigger common data warehouses for statistical production. This means that it is not easy to track the changes made to a certain variable used by a certain Statistical Program. The same people are no longer in charge of all the production phases of a certain statistical product, as in the traditional era. Today, especially researchers are demanding descriptions concerning what has happened to the data and individual variables during the statistical production process. Also, the General Data Protection Regulation (GDPR) is creating pressure for metadata to enable the automation of personal data requests concerning the right to access (Reform of EU data protection rules, 2018). This requires process automation including creating specific logs and metrics, as well as tracking certain data objects with the help of metadata.

It is important to realise that the above described three new roles for metadata are linked to each other. The well-planned metadata for automated processes can most probably be used in a machine-readable way to track the changes made to the data. The discoverability nature of metadata enables us to understand semantically the data we have in different phases of our production, which is a pre-requisite to drive processes.

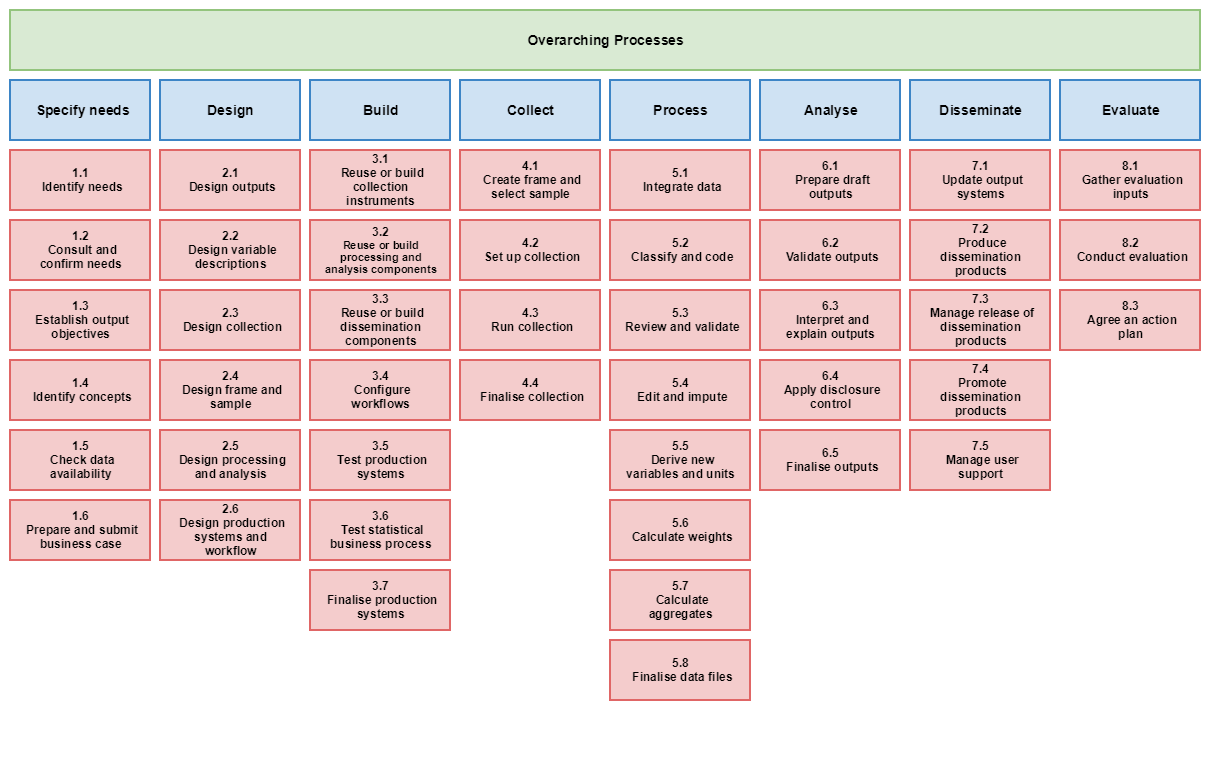
How to handle all three new challenging roles at the same time? The most likely answer is focusing on designing and upper level architectural plans, which can then be followed in all the projects implementing GSIM-based metadata solutions. If we focus on designing, we could avoid ending up in possessing data swamps where nothing can be easily and reliably found and most likely not utilised in common processes and services*.*

Furthermore, Data management policies and technologies will continue to develop. What will be the possibilities to collect, process and analyse future data? How will these be supported from the metadata side? One way of thinking is to create solutions for an NSO aligned with Data Virtualisation best practices (Statistics Estonia, 2018).

**4.** **How these new metadata roles are visible in international standards?**

The statistical production process is described in a reference model called the Generic Statistical Business Process Model or GSBPM. The model was updated to version 5.1 in spring 2019.

**6 Three roles of metadata within the GSBPM framework**



Usability support and Trackability support

Drive processes and Usability support

Tracka-bility support

Usability support

Source: United Nations Economic Commission, 2019b (edited by authors)

In these three roles, usability support, drive processes and trackability support, different phases of the GSBPM are in focus. In some ways, these roles are present in all production phases. Nevertheless, the illustration above addresses the main focus of each phase or set of phases. In some ways, the GSBPM is on a quite high level to map metadata into specific processes. In order to map processes more tightly with metadata roles presented earlier, it would perhaps require expanding the GSBPM one level down, the way e.g. Statistics Canada has planned (Blackwell and Kalonji, 2018).

Another already mentioned UNECE standard in this space is the GSIM (United Nations Economic Commission, 2019a). The Generic Statistical Information Model describes and defines the information objects needed in statistical production, e.g. the GSBPM, and the relationships between those objects. We used the GSIM as a reference to create a matrix (Table 1) to illustrate which object supports what role. The main idea was to show which object prominently supports the selected three roles.

**Table 1. The metadata roles in GSIM framework**



Summing up, we can see that the GSIM Structures part (yellow) supports mostly process drivability. The Concepts part (green) supports interoperability and discoverability thereby offering usability support. Trackability is not represented that much in the GSIM, but the expanded Base part (orange) has some objects that support this role.

In addition to these UNECE standards, useful in this area are also VTL (SDMX, 2018b), as well as SDMX (SDMX, 2011), DDI (DDI Alliance, 2018) and ISO11179. Traditional metadata models SDMX, DDI and ISO 11179 support interoperability well and they are mapped across. VTL is a newly developed model which enables documentation of data transformation and the role to drive processes and traceability support. The GSIM has brought more power to all these implementation models since they are all mapped to the GSIM.

**5. Development projects developing Metadata management in Statistics Finland**

In practice, we have created our own information model, called TIMO, based on the GSIM. Whereas the GSIM is a conceptual model, our information model can be seen as something in between a logical and a conceptual information model. In addition to supporting the usability of metadata, the model is planned to be used as a model against which all objects can be evaluated and added into our Identifier Service. The TIMO-model, that has been developed stepwise to serve the needs of data management, is in very early stages compared with the GSIM. The current version 0.5 mainly includes objects from the green Concept part and from the yellow Structure part of GSIM.

To support automation, i.e. metadata driven production, we have been developing our metadata system in one project related to social statistics at Statistics Finland. STIINA-development is running from 2015 to 2020 and aims to renew the production process of around 70 Statistical Programs inside the social statistics domain. In order to do this, we needed to design a new metadata system called METSY. The new system is a Cloud native system, based on serverless architecture. METSY is GSIM-based and implements several objects of the GSIM Concept part and also some of the Structure part. One of the main ideas has been to build METSY to support services that could be then shared across Statistics Finland. So far, we have built services for data collection and also an editing service using VTL-language. The plan is to create more services and make them compatible with METSY.

Already before METSY, the first GSIM implementation was around 2015 when Statistics Finland made the decision of adopting the GSIM Statistical Classification Model for the new Classification System (Kaukonen et al, 2018). It has been easy to connect this system with METSY since both systems are GSIM-based.

While we have made substantial progress in terms of supporting automation, we have made few advances in supporting trackability. Despite advancing in some parts of the production process there are still many manual processes in between the few automated ones. And some of these processes have proven to be difficult to automate. Full automation is perhaps not a realistic target and a perfect seamless lineage, consequently, might be impossible to reach.

At the same time, we are struggling to find appropriate opportunities to support usability, especially discoverability and interoperability utilising GSIM-based metadata. The relationships between Instance Variables – Representative Variables and Variables play a crucial role here, not to forget the relationship between Concepts and on the other hand Unit Type, Universe and Variable. There is no easy way out but, perhaps, an easier path could be to first focus only on one of the metadata roles stated in this paper. Theoretically, a logical order would be to build up systems to support discoverability first, automation second and trackability third.

**6. Conclusion**

To conclude, we think there are some ways forward that we would like to bring up:

1. Metadata development should have a more strategic focus. Recognising metadata challenges and solutions in Enterprise Architecture context is one way forward. The outcomes, for instance metadata capabilities, could be shared among the Nordic countries.
2. It is crucial to be able to link metadata to services running the production process to realise the benefits of coherently managed metadata. Services use metadata as a fuel to run. We could encourage the development of services that would be documented using standards. This would give real examples of how information could move coherently inside the production process. In the statistical world, at least in a Nordic context, we could then share the descriptions of these standard-based services, in addition to, or instead of, the actual built services (CSPA services).
3. Working with international standards development is a key to understanding them and learning from other NSOs’ best practices in data and metadata management. For example, identifying the GSIM objects in the GSBPM sub-processes could be a revelation to the NSOs trying to implement these models. However, it is still a link between two conceptual models. It gives a lot of reference for implementation, but many issues are probably still left unsolved because the issues are too detailed for conceptual models. Currently, the UNECE Modernisation Group Supporting Standards has several Task teams working around standards. One of them is the GSBPM-GSIM linking Task team, that is working with the issue of linking the two standards together. The Nordic countries are represented in the task team by Finnish and Swedish participants.
4. The Nordic Metadata Meeting has been organised annually since 2013. This has been a useful forum to share ideas and thoughts amongst the so called “metadata people”. If we aim at producing services for the whole NSO and for all the production phases, it is important to expose other experts to these international standards. Therefore, other experts than just metadata developers should be invited to this meeting. The agenda of the meeting could include issues related to these new roles of metadata. The event could even be more of a “Modernizing of Nordic statistics” type of activity including parallel sessions for specific topics.

**7. References**

Blackwell J. and Kalonji F. (2018), Statistics Canada Using Standards, Presentation to ModernStats World Workshop, Available at: <https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2018/mtg1/Canada_BLACKWELL_Presentation.pdf> slide 15 (Accessed: 13 June 2019)

Dalton P. Dunne J. Kelly D. (2015), Modernisation: Evolution or Revolution, Global Conference on a Transformative Agenda 15-16 January 2015 New York Session 2: Innovations in data collection, data dissemination, data access and data analytics, Available at: <https://statswiki.unece.org/pages/viewpage.action?pageId=128451434&preview=/128451434/128516188/Session2_paper_Modernisation%20Evolution%20or%20revolution.pdf> (Accessed: 13 June 2019)

DDI Alliance (2018), Documentation, Available at: <https://www.ddialliance.org/explore-documentation> (Accessed: 13 June 2019)

Information Terms Glossary (2017), Available at: <http://finto.fi/tt/fi/page/?clang=en&uri=t41> (Accessed: 3 May 2019)

Kaukonen E. Davis D. Leino J. Pasila A. Räikkönen T. Saloila M. and Saranpää T. (2018), GSIM implementation at Statistics Finland, Available at: <https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2018/mtg1/Finland_KAUKONEN_Paper.pdf> (Accessed: 6 June 2019)

Population Register Centre (2019), Interoperability platform and method, Available at: <https://yhteentoimiva.suomi.fi/en/> (Accessed: 13 June 2019)

Reform of EU data protection rules (2018), Available at: <https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en> (Accessed: 6 June 2019)

Rouhuvirta H. and Lehtinen H. (2003), Common Structure

of Statistical Information (CoSSI) Definition Descriptions Available at: <http://www.stat.fi/org/tut/dthemes/drafts/cossi_definition_descriptions_v_09_2003.pdf> (Accessed: 3 May 2019)

SDMX (2018a), Glossary, version 2.0 Available at: <https://sdmx.org/wp-content/uploads/SDMX_Glossary_Version_2_0_October_2018.docx> (Accessed: 6 June 2019)

SDMX (2018b), Validation and Transformation Language (VTL), Available at: <https://sdmx.org/?page_id=5096> (Accessed: 13 June 2019)

SDMX (2011) SDMX Standards section 2 Information model: UML conceoptual design, version 2.1, Available at: <https://sdmx.org/wp-content/uploads/SDMX_2-1-1_SECTION_2_InformationModel_201108.pdf> (Accessed: 13 June 2019)

Statistisc Estonia (2018), Development Plan of Statistics Estonia 2018–2022 Statistics Estonia as coordinator of state data governance Tallinn 2018 CONTENTS Introduction, Available at: [*https://www.stat.ee/dokumendid/1231827*](https://www.stat.ee/dokumendid/1231827)(Accessed: 13 June 2019)

Statistics Finland (2014), History of Official Statistics of Finland, Available at: <http://www.stat.fi/meta/svt/taustaa_en.html> (Accessed: 13 June 2019)

United Nations Economic Commission (2009), Common Metadata Framework Part A: Statistical Metadata in a Corporate Context: A guide for managers, Available at <https://statswiki.unece.org/display/hlgbas/Part+A+-+Statistical+Metadata+in+a+Corporate+Context?preview=/8683564/31260677/CMF_Part_A_final%20for%20web.pdf> (Accessed 6 June 2019).

United Nations Economic Commission (2019a), Generic Statistical Information Model Available at: <https://statswiki.unece.org/display/clickablegsim/Clickable+GSIM+v1.2> (Accessed 13 May 2019).

United Nations Economic Commission (2019b), Generic Statistical Business Process Model, Available at: <https://statswiki.unece.org/display/GSBPM/Clickable+GSBPM+v5.1>   
(Accessed 13 May 2019).

United Nations Economic Commission (2019c), Common Statistical Data Architecture, Available at: <https://statswiki.unece.org/display/DA/Data+Architecture+Home> (Accessed 13 May 2019).