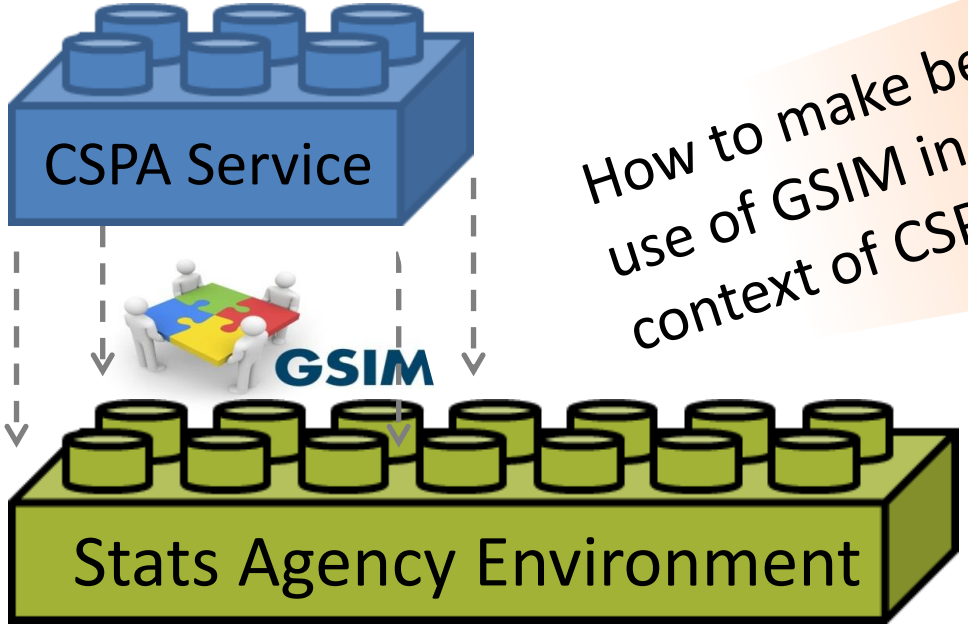
Summary of GSIM Statistical Classifications (made for CSPA LIM)

Klas Blomqvist

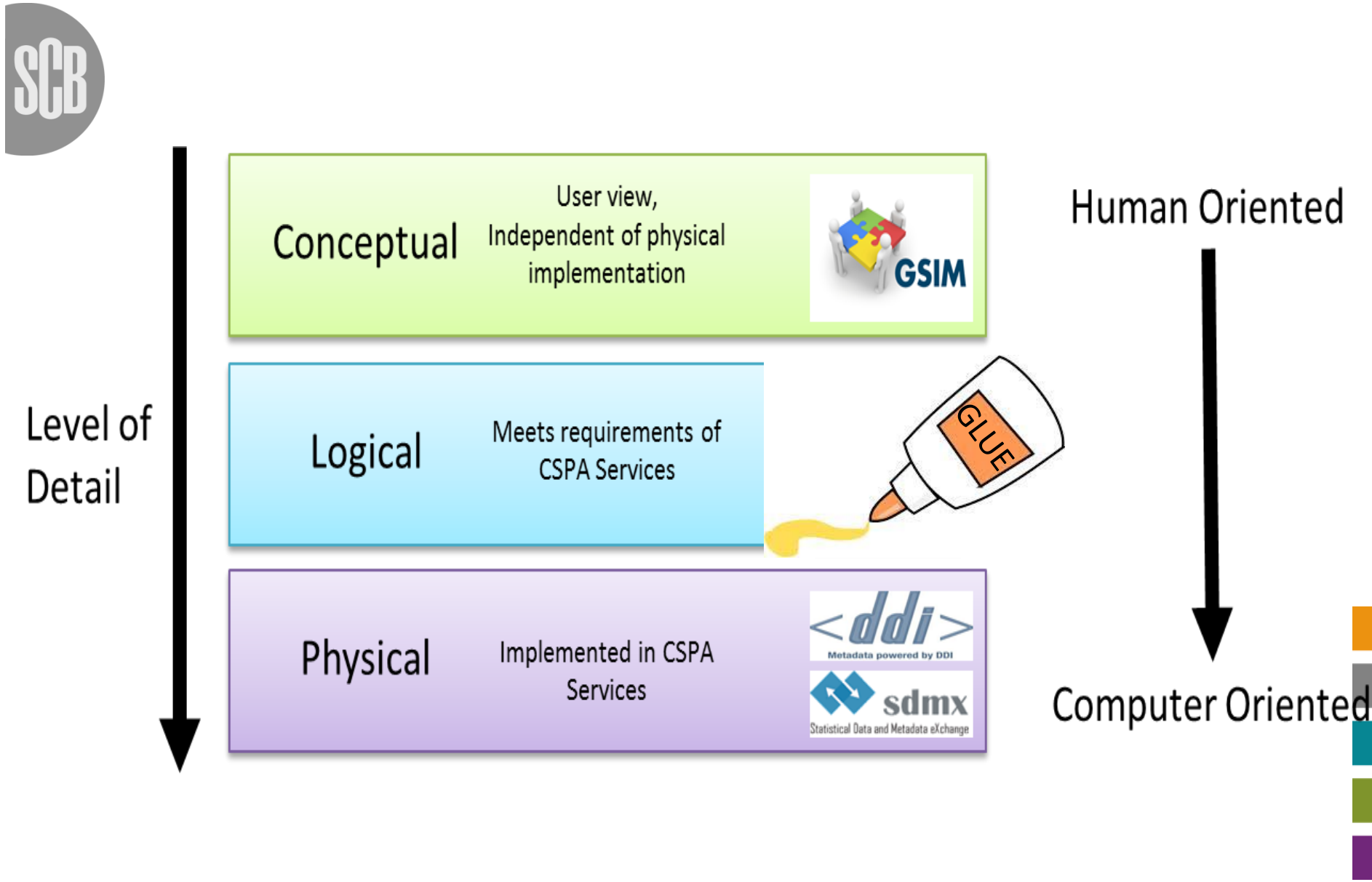
Statistics

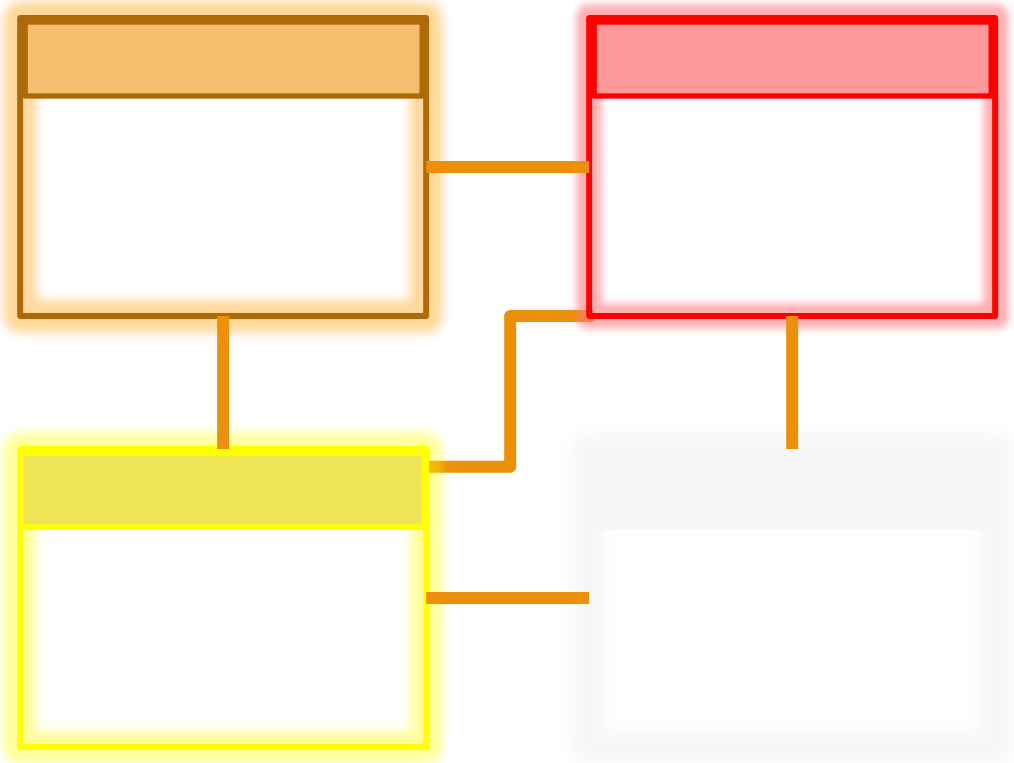
Why LIM

- The gap between the conceptual nature of the GSIM and the practical implementation focus of the CSPA was too wide.
- To bridge this gap, a new layer, a Logical Information Model (LIM), was needed.



How to make best use of GSIM in the context of CSPA


Level of Detail
Conceptual
User view, Independent of physical implementation
GSIM
Logical
Meets requirements of CSPA Services
GLUE
Physical
Implemented in CSPA Services
<ddi>
Metadata powered by DDI
sdmx
Statistical Data and Metadata eXchange
Human Oriented
Computer Oriented



CSPA Logical Information Model

LIM

- The aim of the LIM is to translate the conceptual GSIM information objects into physical specifications of the information that flows in and out of statistical services.
- LIM describes the information objects and logical relationships required to support a CSPA service, in a manner which is consistent with GSIM.
- LIM is independent of the terminology used in existing standards such as SDMX and DDI.

Scope for LIM

- Not all GSIM information objects will make it to LIM.
- The LIM is only concerned with the service and as such will not be taking all GSIM information objects down to LIM level.

Further development of LIM

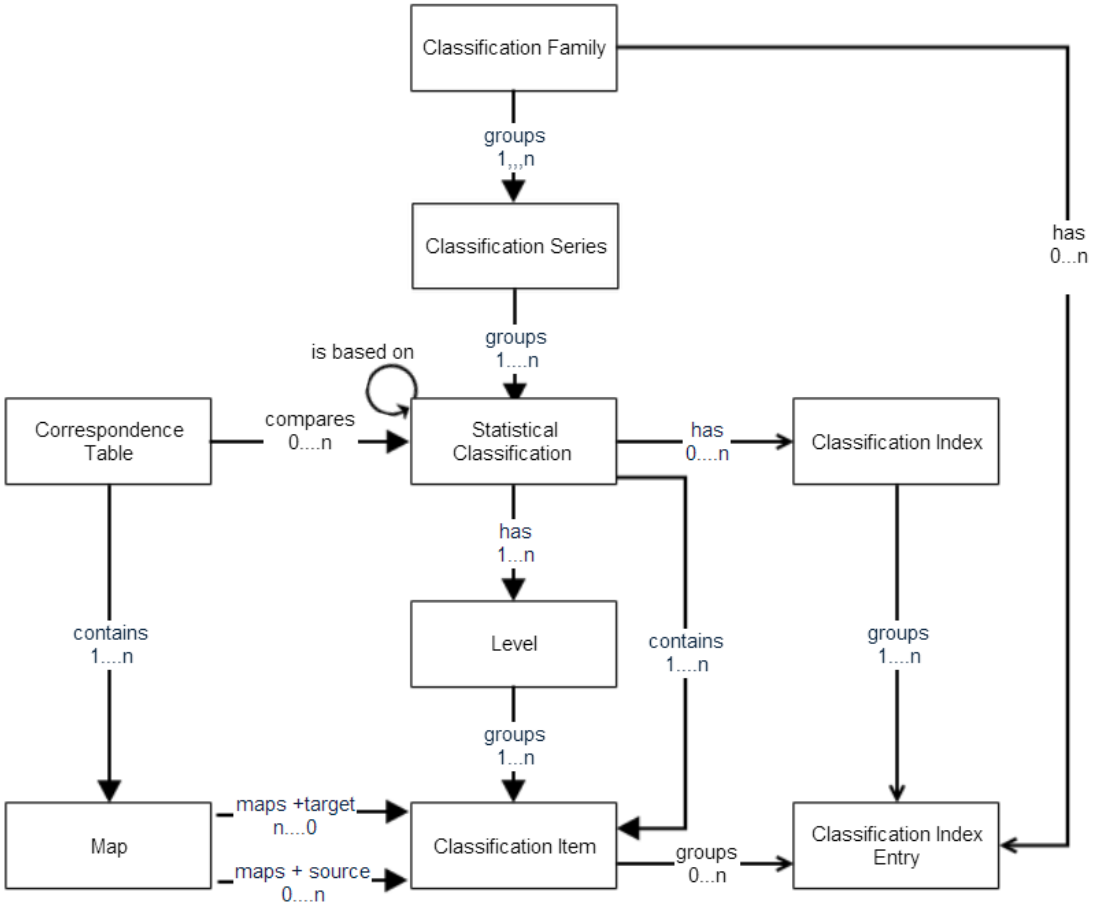
- CSPA Implemention group
- Subgroup LIM
- Variables identified as high priority
  - Help avoid risk of implementors using different Variable incorrectly -> easier to share services
  - Neuchâtel terminology model definitions

GSIM Statistical Classifications Model

GSIM Statistical Classifications Model

- Based on Neuchâtel terminology v2.1.
- Developed by a group of 19 members from 13 different national and international organisations provide common language and common perception of the structure of statistical classifications and the links between them.
- It is both a terminology and a conceptual model.
- It defines the key concepts that are relevant to structuring Statistical Classification metadata and provides the conceptual framework for the development of a Statistical Classification management system.
- Two level structure
  - object types (e.g. Statistical Classification, Classification Item)
  - second level, the attributes associated with each object type

Conceptual model



Overview

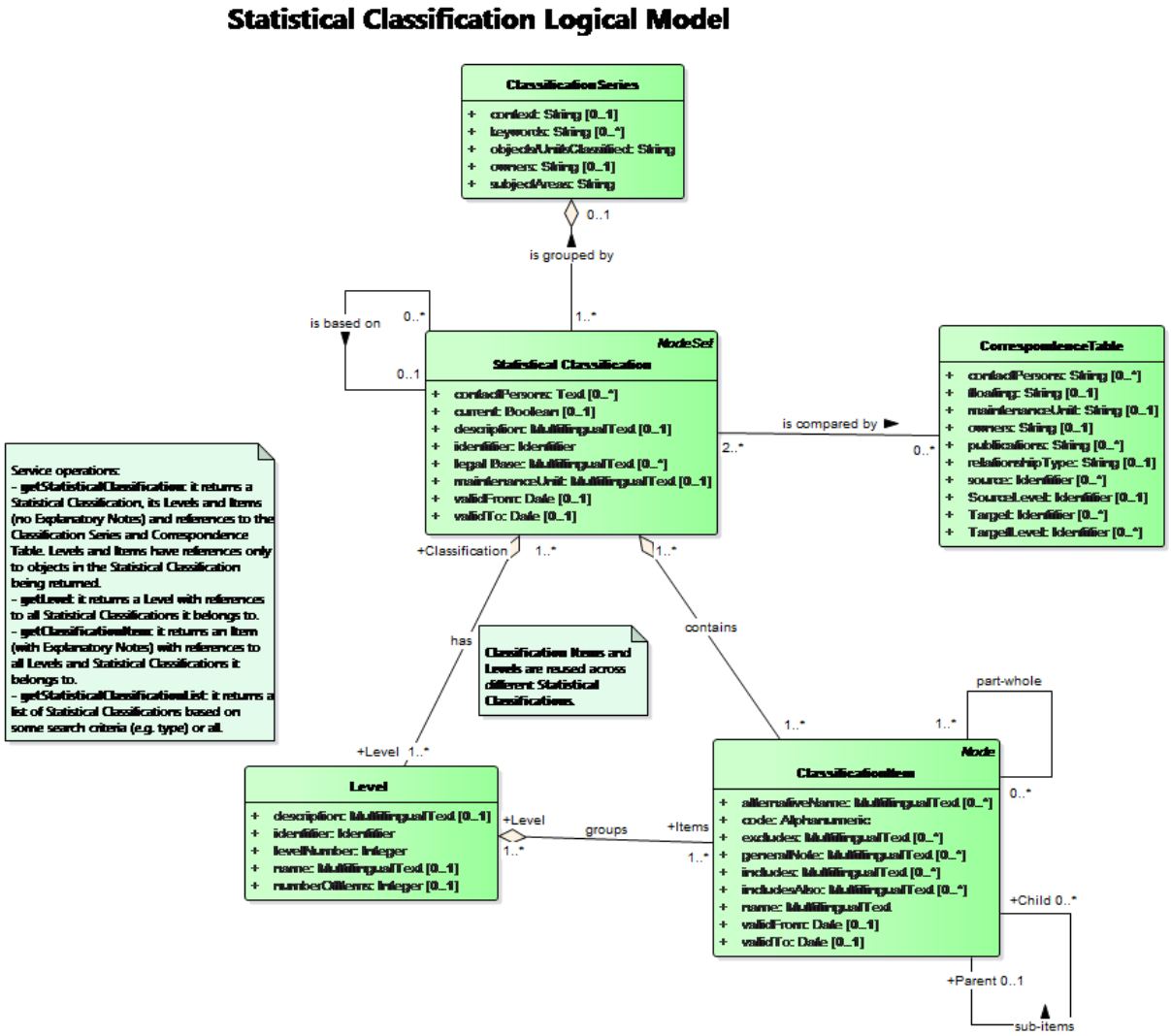
- Classification family (activity classifications)
- Classification series (ISIC)
- Statistical classification (ISIC rev 4)
- Level (Section, Division, Group, Class)
- Classification Item (0311- Marine fishing)
- Correspondence table
- Map (0311 - 03.11)
- Classifiaction index (coding tool)
- Classification index entry

LIM for classifications

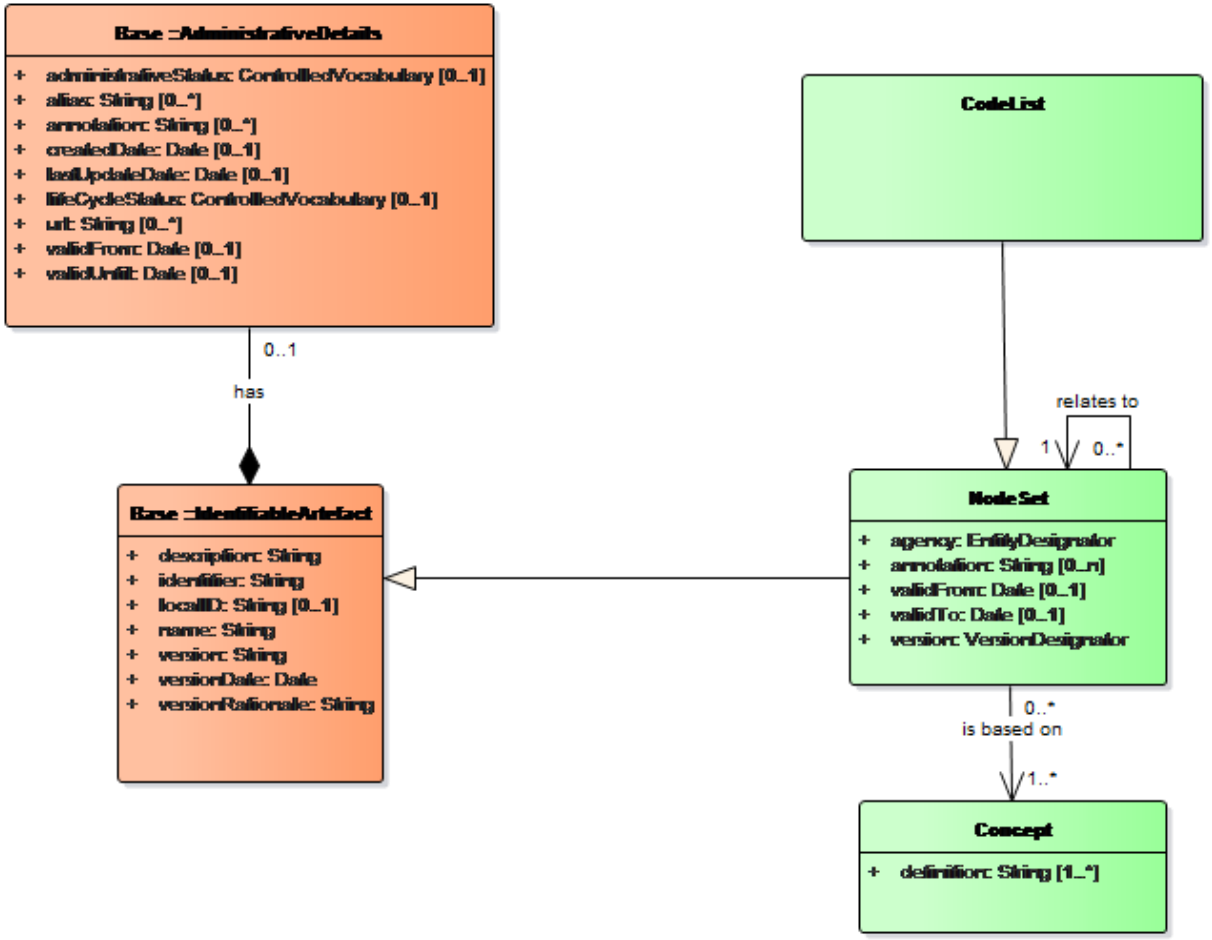
LIM for classifications

- LIM is aimed at service developers. In order to produce CSPA *Statistical Classification* services, seemingly redundant attributes have been removed i.e. replaced by relationships, and the number of attributes has been reduced to cover only those used by more than 60% of those organisations that have documented their use of GSIM *Statistical Classifications* in their GSIM Case studies (August 2015).

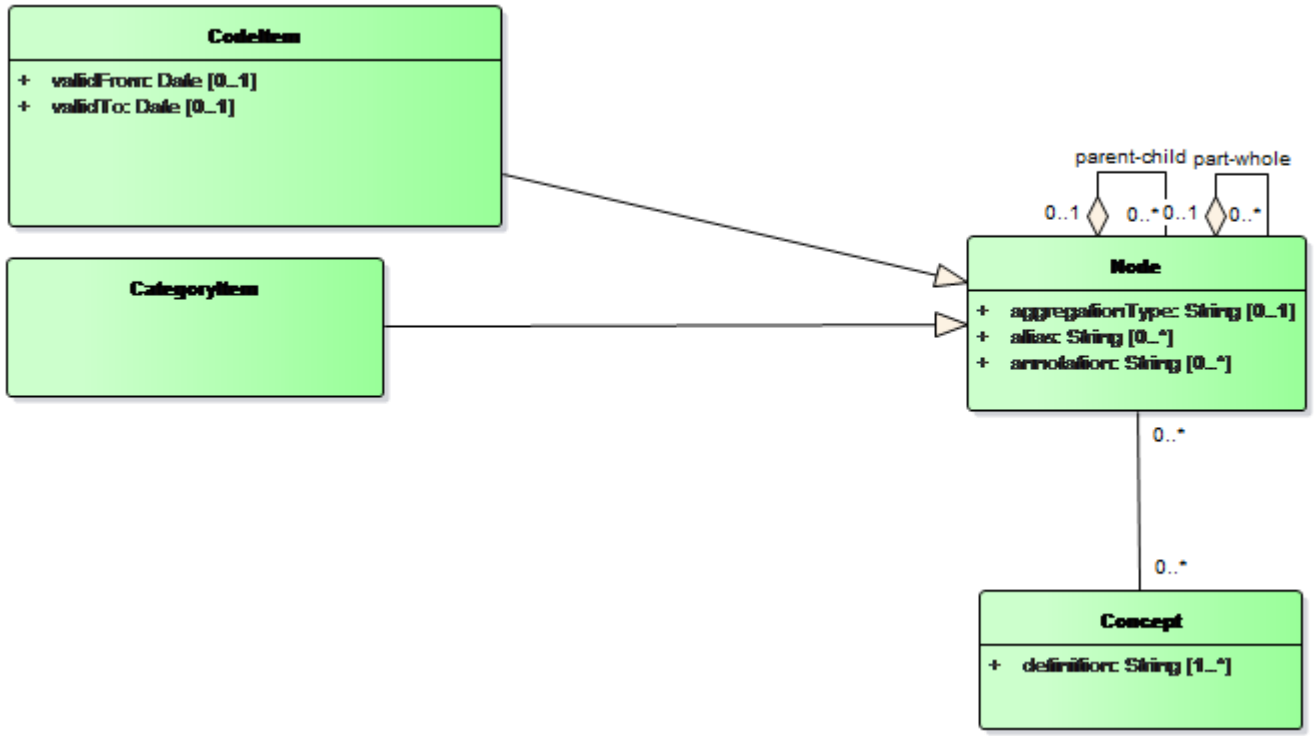
Statistical classification logical model



Code list


Base ::AdministrativeDetails
+ administrativeStatus: ControlledVocabulary [0..1]
+ alias: String [0..*]
+ annotation: String [0..*]
+ createdDate: Date [0..1]
+ lastUpdateDate: Date [0..1]
+ lifeCycleStatus: ControlledVocabulary [0..1]
+ url: String [0..*]
+ validFrom: Date [0..1]
+ validUntil: Date [0..1]
0..1
has
CodeList
relates to
1
0..*
Base ::IdentifiableArtefact
+ description: String
+ identifier: String
+ localID: String [0..1]
+ name: String
+ version: String
+ versionDate: Date
+ versionRationale: String
NodeSet
+ agency: EntityDesignator
+ annotation: String [0..n]
+ validFrom: Date [0..1]
+ validTo: Date [0..1]
+ version: VersionDesignator
0..*
is based on
1..*
Concept
+ definition: String [1..*]

Code item


CodeItem
+ validFrom: Date [0..1]
+ validTo: Date [0..1]
CategoryItem
parent-child
part-whole
0..1
0..*
0..1
0..*
Node
+ aggregationType: String [0..1]
+ alias: String [0..*]
+ annotation: String [0..*]
0..*
0..*
Concept
+ definition: String [1..*]

Thank you…

CSPA - Logical Information Model (LIM)

- GSIM provides a common language for describing the information objects relevant to the statistical production process
- Having a common language increases the ability to compare information within and between statistical organizations.
- The Common Statistical Production Architecture (CSPA) is a reference architecture for the official statistics industry
- There is a gap between the conceptual nature of the GSIM and the practical implementation focus of the CSPA was too wide. To bridge this gap, a new layer, a Logical Information Model (LIM), was needed. The aim of the LIM is to help developers of CSPA-compliant services by translating the conceptual GSIM information objects into physical specifications of the information that flows in and out of statistical services.

- 1. The Generic Statistical Information Model (GSIM) provides a common language for describing the information objects relevant to the statistical production process. Having a common language increases the ability to compare information within and between statistical organizations. The Common Statistical Production Architecture (CSPA) is a reference architecture for the official statistics industry, providing the blueprint for designing and building statistical services in a way that facilitates sharing and easy integration into statistical production processes within or between statistical organisations
- 2. At a sprint session held in Ottawa in February 2015, participants concluded that the gap between the conceptual nature of the GSIM and the practical implementation focus of the CSPA was too wide. To bridge this gap, a new layer, a Logical Information Model (LIM), was needed. The aim of the LIM is to help developers of CSPA-compliant services by translating the conceptual GSIM information objects into physical specifications of the information that flows in and out of statistical services.
- 3. LIM refines the conceptual definitions from GSIM, and describes the information objects and logical relationships required to support a CSPA service, in a manner which is consistent with GSIM and independent of the terminology used in existing standards such as SDMX (Statistical Data and Metadata eXchange)
- [1]and DDI (Data Documentation Initiative)
- [2]. It supports consistent use of SDMX, DDI and other implementation standards in reusable CPSA services, whilst also making it easier for any organisations that do not use SDMX or DDI to implement CSPA-compliant services.
- 4. One of the requirements for a CSPA Service to be included in the CSPA Statistical Service Catalogue is that it is compliant with the architecture and it is specified using the LIM.
- 5. A pragmatic approach has been followed during the development of the LIM. Only those parts actually needed by CSPA Services during 2015 being developed during that year. Other parts will be worked on as and when needed. The LIM is designed in accordance with the LIM Design Principles (outlined in Annex1).The scope of the LIM, and the current state of development at the end of 2015 are illustrated in the following table. The remainder of this document describes the parts of LIM that have been developed so far

Using LIM

- **Required use of LIM in the specification of a CSPA Service**
- 6. The LIM is used in the specification of the service. At this stage of design, the service will already have an approved Statistical Service Definition and the service will have been defined in terms of GSIM and the business process flow.
- 7. For a CSPA service to be considered compliant with the architecture and approved for inclusion in the CSPA Statistical Service Catalogue, it is necessary for the Service designer to explain what the expected interface is. The CSPA Service Specification template requires the data and the process logic and methods to be modelled in LIM.
- 8. To do this, the Service designer must identify the relevant LIM objects for the service interface. There are some objects that are likely to always be needed (for example process design, data structures). For each service, some lower level details may be required to determine the particular structure of some of the objects identified from LIM (for example the detailed structure of an identified *Code List*). The service design should identify the specific behaviours of the interface (i.e. would you return a dataset, or a reference to a dataset.) Knowing the architectural patterns to be used and the communication platform may affect which LIM objects are used in the interface.
- 9. Service Designers can use the Process Package to describe the methodology to be implemented by the Service Builder. There are a number of objects within LIM that are used to describe processes. These objects range from those used to design the Business Process, including the methods and rules used to define the actions encompassed in the process, through to the objects needed to encapsulate the execution-time processing, identifying the inputs used and the outputs produced at each step.
- 10. For a CSPA Service, these objects are useful tools to enable the complete description of the Service, the actions the Service will undertake and the inputs and outputs needed by the Service to complete these actions. These are documented in the Service Specification.
- 11. The distinction between the design-time and execution-time process objects indicates a focal point for the Service Designers and Service Builders. While this split isn't mutually exclusive between the two roles, generally speaking a Service Designer could develop a service using the LIM objects for describing a process, while the Service Builder would focus on the execution-time objects.
- 12. For example, in the Service Specification document used to describe the CSPA Service, the *Process Step* objects would be used to encapsulate the information about what the Service will do. For the Service Builder, the *Process Input*s and Outputs identify what information is to be passed across the service boundaries for the service to perform its function (i.e. the service interface).
- 13. The service team will need to validate their proposal with the CSPA Implementation Group. In the case where the LIM does not include sufficient objects or attributes, this group can develop the model further to meet the needs of services. Once approved the Service Builder will take this specification and implement it based on the standard(s) that will be used.
- **Other ways to use LIM**
- 14. Service Builders can use LIM to describe any (machine actionable) orchestration that is encapsulated by the Service. This will help others to understand the data models and process logic in the service (so they can use it) and reuse existing implementations of LIM and their physical representations. They can record these LIM/Process diagrams as appropriate.
- 15. Assemblers need to know the LIM objects so they can understand whether they can implement the service. They also need to understand the interface, so that they can determine if the service will be able to be integrated.

- 6. 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. The definition of the LIM and the specification of physical representations based on this logical model provide the way to translate agreed GSIM information objects into consistent, standards aligned, physical inputs and outputs for CSPA services.
- 17. For the Service Specification, we need LIM to describe information objects and the precise logical relationships between them in a manner which is consistent with GSIM. The primary interest for 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. 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.
- 18. 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. The following principles guide this decision.
- If there is a clear and unambiguous influence from an object in the logical model of another standard (such as DDI or SDMX) then:
  - If that object has mandatory attributes, those should be mandatory in the LIM or the LIM must provide guidance to default that attribute once implemented.
  - If that object has a relationship with another object that significantly affects its behaviour or use, that same relationship must exist in the LIM.
  - Future known changes to that object should be accommodated for in the LIM.
- Try to maintain alignment with equivalent objects in the supporting physical standards, developers will need this alignment in place in order to implement. This will also maximize the ability for developers to reuse available toolkits.
- Only 'improve' objects from underlying physical standards once you have exhausted options to have those changes implemented in the physical standard; if the physical standard is changed, then reflect that change; if the physical standard isn't changed, explain the reason for the improvement.
- Consider emerging and alternative standards when trying to identify the need for new objects in the LIM – other standards may have already expended the effort required to address the use case.
- Consider using existing mapping work (for example between DDI and SDMX,) to help align these standards with the LIM.
- Make efforts to be aware of proposed and agreed future changes to key physical standards like DDI and SDMX.

GSIM Statistical Classifications Model

- **Generic Statistical Information Model (GSIM) Statistical Classifications Model**
- 3. In June 1999, a meeting on terminology was held in Neuchâtel, Switzerland, with participants from the statistical offices of Denmark, Norway, Sweden and Switzerland and the software developers in Run SoftwareWerkstatt. This was the start ofthe "Neuchâtel group". The aim of the group was to clarify some basic concepts and to arrive at a common terminology for classifications. The terminology defined the key concepts that were relevant for how to structure classification metadata andprovided the conceptual framework for the development of a classification database. The work listed and described the typical object types of a classification database, and the attributes connected with each object type.
  4. The development of the model had a practical focus as all of the participating National Statistical Organisations (NSOs) planned to use it in their own implementation of a classification database. The most important purposes for developing aclassification database were:
- to make accessibility and maintenance of classifications easier, and
- to ensure common use of classifications across different fields of statistics.
  A central database was the preferred solution because it realised one of the important principles of metadata - document and update once (centrally), and reuse wherever it is relevant. The Neuchâtel terminology model: Classification databaseobject types and their attributes (version 2.0) was released in 2002.
  5. Later, Statistics Netherlands joined the Neuchâtel group, and a new version of the terminology, version 2.1 (with one new object and one new attribute), was released in 2004.
  6. It was essential for the Neuchâtel group that the terminology should be flexible and independent of IT software and platforms. This resulted in different classification database implementations for the participating NSOs, according to specificneeds and policies. Also,
  it was always an important premise for the group that the work should be public and available to anyone free of charge.
- 7. Many countries have at least partially implemented the model
- [1]. After years of practical experience, several of the implementing countries expressed a desire to see some revisions to the model. As the Neuchâtel group no longer existed, a possible revision was discussed at the 2011 METIS Workshop [2]. Subsequent to the workshop, the METIS Steering Group contacted the UN Expert Group on International Statistical Classifications to work on the revision of the Neuchâtel model. As a result, a joint working group was created, bringing together classification and statistical metadata experts.
- 8. At the same time, a project sponsored by the High Level Group for the Modernization of Statistical Production and Services was reviewing the Generic Statistical Information Model (GSIM)
- [3]. GSIM provides the information object framework supporting all statistical production processes such as those described in the Generic Statistical Business Process Model (GSBPM)
- [4], giving the information objects agreed names, defining them, specifying their essential properties, and indicating their relationships with other information objects. In the development of GSIM, the objects related to classifications were mostly drawn from the Neuchâtel Terminology Model.
- 9. During the revision work it was discussed and decided that for the future the Neuchâtel model for classifications will be part of GSIM. Several objects and attributes have been changed during the revision process, and the revised model will inpractice be an annex to GSIM.