منظمة الأغذية والزراعة للأمم المتحدة | 联合国粮食及农业组织 | Food and Agriculture Organization of the United Nations | Organisation des Nations Unies pour l'alimentation et l'agriculture | Organización de las Naciones Unidas para la Agricultura y la Alimentación

COORDINATING WORKING PARTY ON FISHERY STATISTICS

Intersessional Fishery Subject Group Meeting

Swakopmund, Namibia 25-27 February 2015

Concept note on business Metadata

Author: FAO/Eurostat/OECD

Fisheries business Metadata and the CWP role for global harmonization

FAO - Erik van Ingen - CWP
Swakopmund, Namibia, 25-27 February 2015

Introduction

This concept note intends to briefly discuss the *Information Standards* which are relevant for Fisheries and therefore for CWP.

Information Standards are also referred to as metadata standards, see also this useful Wikipediapage http://en.wikipedia.org/wiki/Metadata_standards

From here on, terms written in **bold** in the introduction, are chapters in this document.

Why are *Information Standards* important for a business? Read **CWP and information standards.**

Information standards do enable exchange of information. It is therefore relevant in this document to discuss also terms like **Webservices, RESTful and API,** which are heavily used in modern IT.

Institutes, governments, businesses do work with information, there are many different processes of information. This can be captured in process models. Therefore the **GSBPM process model** is touched upon in this document.

Information standards are sometimes very specific for a certain process. Think of the processes of the collection of position data of vessels (VMS) or species observation data (Occurrence records). Another very important process is the Collection and dissemination of statistical data. Read **Observation and statistical data** to understand why this document focuses on statistical data.

What is a schema based information standard and how does this relate to non domain specific information standards? See **Schema, Domain and Artefact based.**

Information standards do enable the Harmonization of information on a global level.

The last couple of years, there is a lot of hype around masterdata. See also Gartner on this:

http://www.gartner.com/it-glossary/master-data-management-mdm and read **Masterdata and information Standards.**

Which *Information Standards* are relevant for Fisheries? Read **Relevant Information Standards for Fisheries.**

The document ends with For CWP consideration and discussion.

CWP and information standards

The main functions of the CWP are described in this page http://www.fao.org/fishery/cwp/en and in this chapter we will try to put them in the context of information standards.

continually review fishery statistics requirements for research, policy-making and management;

This main function calls upon the CWP to be agile and responsive. The fishery statistics requirements do change with the speed of how the world is changing, relating itself to internet. Internet makes information available at anytime and everywhere. In the context of this main function of the CWP, advanced technologies as of today, like webservices API and RESTful are of vital importance, see therefore also the chapter Webservices, RESTful and API

agree on standard concepts, definitions, classifications and methodologies for the collection and collation of fishery statistics;

To implement this main function, information standard can be adopted and developed as concrete means to materialize agreements on fisheries data and metadata. See therefore also Relevant Information Standards for Fisheries See for agreeing on standards Harmonization of information

make proposals for the coordination and streamlining of statistical activities among relevant intergovernmental organizations.

This function calls upon a process model or framework, which goes beyond only an information standard and over spans institutes. Therefore this document touches upon the GSBPM process model. But it touches also on Harmonization of information

Webservices, RESTful and API

The webservices took off around the year 2000. Webservices enable machines to read from other machines. In other words, a FAO statistical data collection ‘machine’ can read from the country statistical data dissemination ‘machine without human intervention. Webservices took off with SOAP and XML, as of today being replaced by RESTful and JSON. We will not describe these terms further here and more focus on its application for CWP. What is notable

is the rise of Application Programming Interface (API) on the internet. Google, Twitter, Facebook, News Agencies to publish a lot of information through an API. An modern API is often a RESTful webservice, publishing information according an implicit or explicit information standard.

Why are APIs relevant for CWP? Because APIs can be instrumental in all the main functions of the CWP. It would enable the automation of all main functions of the CWP, except for those actions where human intervention is needed and has an added value. This would make the CWP much more attractive and easier and cheaper to adopt for its members.

GSBPM process model

See also
http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

GSBPM stands for Generic Statistical Business Process Model

The GSBPM is developed by national and international statistical institutes, in the METIS initiative from the UNECE. It has reached already version 5 and is extremely valuable as a reference for the design of statistical processes.

Later on the GSIM was developed:
http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

GSIM stands for Generic Statistical Information Model

Can the GSIM be useful for the CWP? Yes, as an reference information model. But then the mapping to an implementation standard needs to be made, see therefore Relevant Information Standards for Fisheries.

Observation and statistical data

What is the difference? Observation data is a measure taken at a certain moment in time. Observation data is sometimes also referred to as raw data or operational data. Statistical data is often processed and aggregated observation data.

By the now the distinction is roughly clear, we do not intend nor pretend in this document to exhaustively define the two terms.

It is clear that the CWP relates to statistical data. The question is whether the CWP needs to relate also to observation data? There are upcoming standards coming up like DDI,

FishFrame and FLUX which merely relate to observation data. Relating to the main functions of the CWP, the suggestion is that CWP also relates to observation data standards, because of its role in research, policy-making and management. The other suggestion is that the CWP could focus on statistical information standards as a first priority, while including observation standards, time permitting.

Either observation or statistical data, the need for codelists seems to be universal. Reference codelists for species, countries, gears and so forth.

Before all was done on paper. IT progresses and and enables electronic collection at field level. This creates new opportunities for setting standards upstream. CWP should acknowledge these developments and reconsider its role and areas of intervention.

Schema, Domain and Artefact based

Information standards are often materialized in a Schema.

A schema can define for instance that fisheries capture time series data has three dimensions, Country, Species, Area and the value is described in tonnes. Doing so, the information standard would be Schema based, holding domain specific information like capture, country, area and tonnes.

An artefact based information standard has an extra layer. The first layer is the Schema, where the notion of a time series data is defined. On artefact layer, an artefact is defined that a specific time series, e.g. capture, deals with Country, Species, Area and Tonnes.

	characteristics	pro	con	examples
schema based	● often domain specific ● 2 layers (schema+data) ● **Single Maintenance Model**	● less freedom ● pragmatic ● easier to implement ● instrumental for the process of central led harmonization	every domain has its own standard	FLUX, FishFrame, DublinCore, FIMES
artefact based	● not domain specific ● 3 layers (schema+artefact +data) ● **Delegated Maintenance Model**	● more freedom ● instrumental for the process of ‘organic’ harmonization	higher complexity	DDI, SDMX

The table mentions Single and Delegated maintenance models.

Single Maintenance Model
The schema based model has only one extra level (besides the data) and that is the Schema. All users of the Schema need to agree on the Schema. To ease this, the Schema can be split in multiple schema with parent and children, where children Schemas can cover subdomains.

Delegated Maintenance Model
The Artefact based model has two extra levels (besides the data). The top level is the Schema and has often a Single maintainer. On artefact level, everybody can implement whatever they like. This model gives a good infrastructure of the development of information standards on national, regional or international level.

Harmonization of information

Harmonization of information is the process of agreeing upon standards and aligning the different standards.

Harmonization can be central led, e.g. there is one entity which dictates the standard. The dictating entity can be anything, a country, region, institute, ec.

Harmonization can be done in a hybrid of central and decentral. In the previous chapter this was referred to as a delegated maintenance model.

Harmonization of information is extremely important in order to make information more readable and predictable. This enables the human consumption of the information and even more the automated consumption (machine readability).

Masterdata and information Standards

See a definition on Master data in the wikipedia: http://en.wikipedia.org/wiki/Master_data

Without stepping into a discussion on its definition, it is clear that the industry is giving great importance to the management of Master data. In that stream, the CWP is therefore discussing the role of information standards, vital in the process of Master data management.

Relevant Information Standards for Fisheries

Relevant standard for Fisheries are UN/CEFACT-FLUX, SDMX, FishFrame, DDI and DarwinCore.

For CWP consideration and discussion

- The CWP can perfectly combine high level tasks with providing guidance when it comes to concrete Fisheries relevant information standards.
- "Moore's law" says that the number of transistors in a integrated circuit doubles approximately every two years. Where IT systems were developed years ago with limited memory, disk space and network bandwidth in mind, today IT systems can go almost wild. Our world is evolving and that the new IT capacities might invite CWP members to realize that a shift of the areas of work of CWP might be necessary.
- FLUX and SDMX artefacts can be published stamped by the CWP as the maintaining agency.
- SDMX artefacts can be published in the SDMX Global Registry, stamped by the CWP as the maintaining agency.
- The work of the CWP will be mapped to the GSBPM and the GSBPM will guide the work of the CWP.
- CWP needs to have a information standard agnostic approach in providing guidance
- Make sure all the concepts & definitions are consistent applied in all the relevant standards.
- Prio2: Time & budget permitting, the CWP can go standard specific (e.g. FLUX&SDMX Catch DSD)
- Recommending participation of RFB agencies in FLUX and SDMX projects
- For domain specific schemas: in reviewing them and ensuring coherence among code lists referred to among these schemas.
- Promoting the structures, concepts and definitions as valuable towards any data exchange and harmonization effort.
- Recommending a MDM initiative for fisheries code lists to facilitate datasets harmonization among the various schemas.

Fisheries business in SDMX

FAO, Erik van Ingen

Swakopmund, Namibia, 25-27 February 2015

2001

Statistical Data and Metadata Exchange

SDMX is an initiative to foster standards for the exchange of statistical information

SDMX

sdmx
Statistical Data and Metadata eXchange

CL = Codelist

(classification, controlled vocabulary, list of codes with description)

DSD = Data Structure Definition

The most valuable SDMX artefacts are CL and DSD

CL & DSD

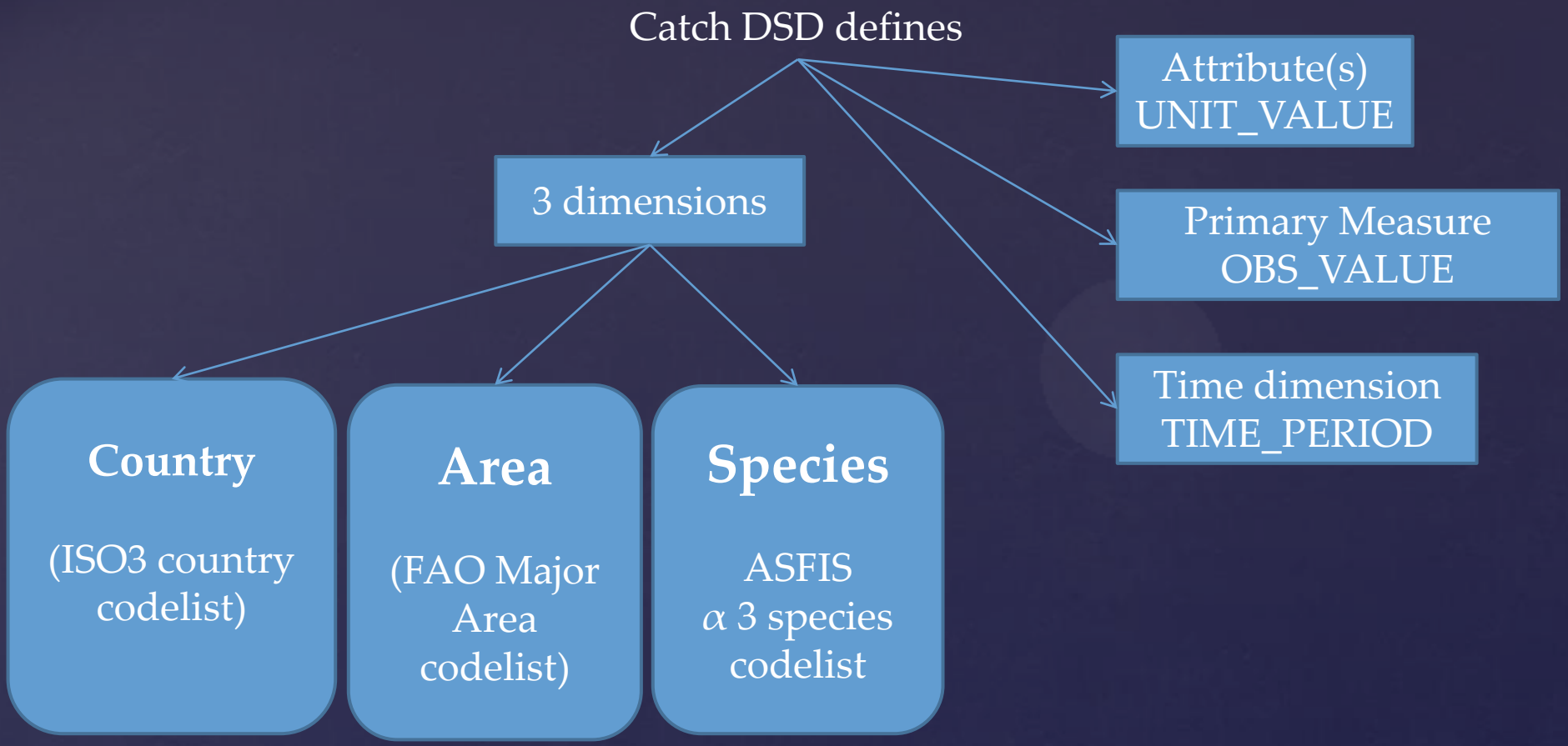
CL & DSD

The DSD defines the dimensions and the related codelists

DSD's and CL's have their own independent life, maintenance and lifecycle

Both DSD and CL have a specified version and maintenance agency

Year	Aquatic Species	Production Area	Country	Quantity [t]	Quantity Unit
2012	Yellowtail flounder	Atlantic, Northwest	Ukraine	0	t
2012	Yellowtail flounder	Atlantic, Northwest	Republic of Korea	0	t
2012	Yellowtail flounder	Atlantic, Northwest	Russian Federation	84	t
2012	Yellowtail rockfish	Pacific, Eastern Central	United States of America	2	t
2012	Yellowtail rockfish	Pacific, Eastern Central	Poland	0	t



SDMX Catch DSD

SDMX Registry

SDMX artefacts are published in a SDMX Registry (web based)

Organisations, countries, regions can decide to set up their own SDMX Registry

SDMX Global Registry

Internationally recognized SDMX artefacts can be promoted to global level

They will be published in the SDMX Global Registry

SDMX Global Registry hosts currently 63 CL and 9 DSD

SDMX in Fisheries

Eurostat has 2 DSDs on Fisheries landing and 3 DSDs on aquaculture

The FAO ASFIS species list and the FAO Major Area list are published in both the FAO and Eurostat SDMX Registries

SEIF

SEIF =

SDMX for Eurostat, ICES and FAO

SEIF has developed the Catch DSD

New version of the Catch DSD is published every year in the Eurostat SDMX Registry

SDMX out there

IMF, OECD and BIS are active on developing DSD collaboratively

Eurostat has a SDMX Registry with 748 CLs and 246 DSDs

African Development Bank and StatCom Africa are implementing SDMX in African countries

SDMX Global Registry has taken off

SDMX has a 2 very active working groups, 1 technical and 1 statistical with each about 20 members

SDMX has never been hyped but always kept on growing steadily, with increased international support

IRD challenges the SEIF Catch DSD as not able to respond to their need

Should they produce yet another DSD?

How such developments of alternative DSDs for each new case would hamper harmonization and interoperability aspects? What is the response there? What could be the role of CWP?

SDMX practical challenges

SDMX practical suggestions

model

- Do understand the importance of data modeling
- Model your data
- Publish your local CLs and DSDs in your SDMX Registry

harmonize

- Relate to partners, discuss your models, CLs and DSDs

standardize

- Develop, design, agree upon a limited amount of CL's and DSD's
- Publish them in the Global SDMX Registry

CWP

verbose – complex - not user friendly

not enough tooling

difficult/expensive to build up SDMX capacity for small countries/organizations

open/closed community

Main identified issues

data harmonization

data collection/dissemination

good international adoption (WorldBank, IMF, FAO, Eurostat, etc)

not domain specific

enables organic growth (3 layer model)

delegated maintenance model

Tangible benefits

Tool	Developed by	Description	Status %
Cotrix	iMarine/FAO	Codelist management (export to the SDMX Registry)	90
Grade/FLOD	FAO	Linked Open Data management /Fisheries Linked Open Data	30
RefPub/RefPlus/ RefVis	FAO	Reference Data Publication/Repository/Visualization	40
SDMX Connectors	Bank of Italy	using SDMX data in statistical packages and tools (EXCEL, R, Matlab, SAS)	100
SDMX Registry Data Structure Wiz SDMX Converter SDMX Visualisation	Eurostat	Tools for working with SDMX standards and guidelines	100
Others	SDMX community	www.sdmxtools.org www.sdmxsource.org	ongoing

Tooling

SDMX enables organic growth, provided some domain specific standard setting bodies (such as CWP) play their part at the artefact level

(I) CWP consideration and discussion

- Instantiate a project for defining SDMX fishery specific artefacts, including resource allocation & funding
- Make sure the presented datamodeling-harmonization-standardization is backup with tooling in iMarine
- Make sure all the SDMX codelists and datastructures are coherent with the FLUX equivalents
- CWP recommends the participation of RFB agencies in projects like FLUX, SDMX and SEIF

(II) CWP consideration and discussion

thank you

FAO, Erik van Ingen

Swakopmund, Namibia, 25-27 February 2015

Fisheries business Metadata & The CWP role for global harmonization

FAO, Erik van Ingen

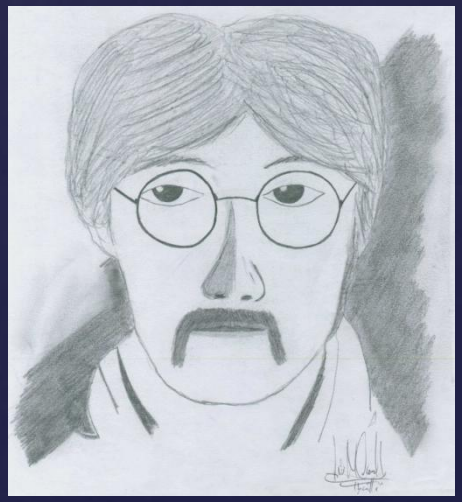
Swakopmund, Namibia, 25-27 February 2015

all the people could

find – use - load – display - interpret - interchange – analyze

all fisheries data

Imagine



Harmonization
Standardization
Interoperability

Better data at lower cost
Fisheries Management

Business Case

FLUX - Fisheries Language for Universal eXchange - A single language for all data exchanges

SDMX - Statistical Data and Metadata eXchange - Foster standards for the exchange of statistical information.

DDI - Data Documentation Initiative - A metadata specification for the social and behavioral sciences

FiMES, DarwinCore, OGC and ISO Observations and Measurements , etc.

Information Standards

Standard setting bodies

standard	body
FLUX	UN/CEFACT
SDMX	BIS - ECB - EUROSTAT - IMF - OECD - UN - World Bank
DDI	Widespread community
FiMES	FIRMS & FAO
DarwinCore	TDWG (Biodiversity Information Standards)
OGC WFS, WMS, WCS	Number of Active Members: 509 WFS accepted as ISO Standard
Observations and Measurements	ISO

Observation data & Statistical data

Observation data (also referred to as raw data or operational data) is a measure taken at a certain moment in time (FLUX)

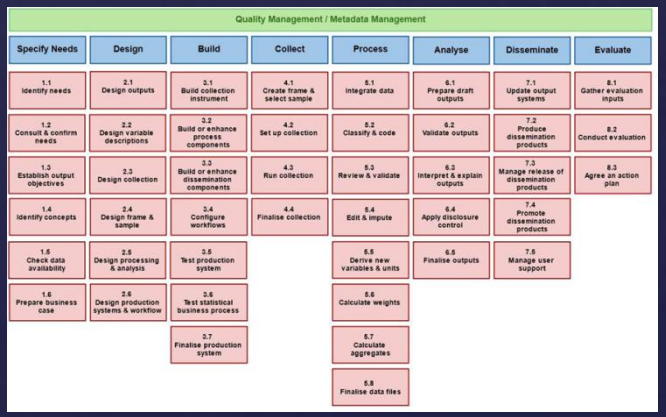
Statistical data (timeseries) is often processed and aggregated observation data (SDMX)

Sample #	pH
1	6.39
2	6.42
3	6.35
4	6.85

Generic Statistical Business Process Model

from United Nations Economic Commission for Europe (UNECE)

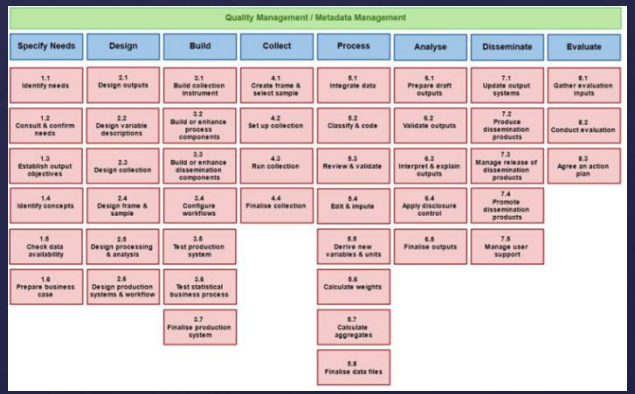
GSBPM



The GSBPM is a reference model for the ‘statistical data process’

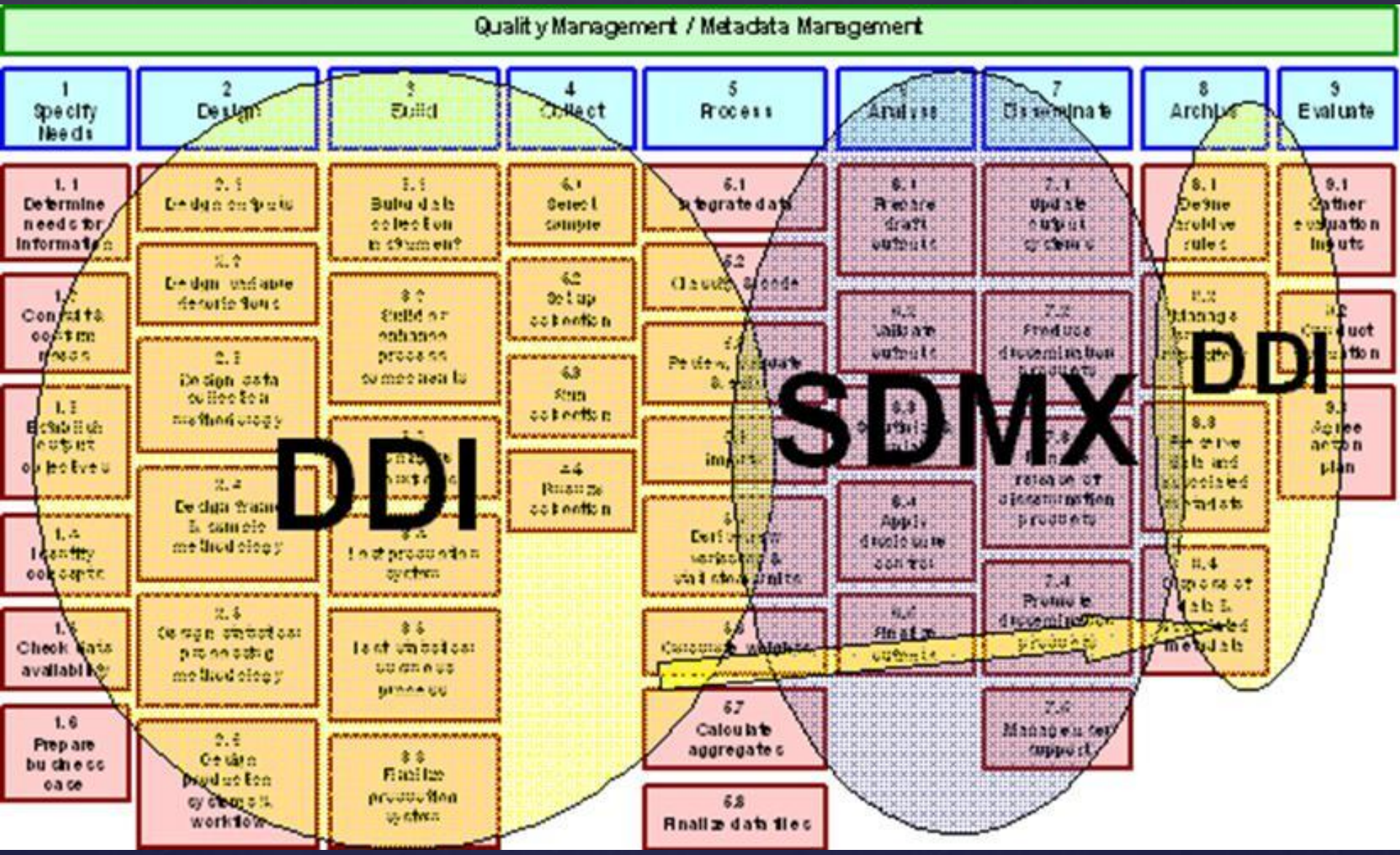
The GSBPM can help to put information standards within a context

GSBPM & CWP



Quality Management / Metadata Management

Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			


Quality Management / Metadata Management
1 Specify Needs
2 Design
3 Build
4 Collect
5 Process
6 Analyse
7 Disseminate
8 Archive
9 Evaluate
DDI
SDMX
DDI

Schema or artefact based

Schema based has 2 layers:

1. schema (domain specific)
2. data

Artefact based has 3 layers:

1. schema
2. artefact (domain specific)
3. data

For the techies in the audience:

- A **schema** is often an XML schema
- An **artefact** of often an XML document
- The **data** is often an XML document, also called a **message**

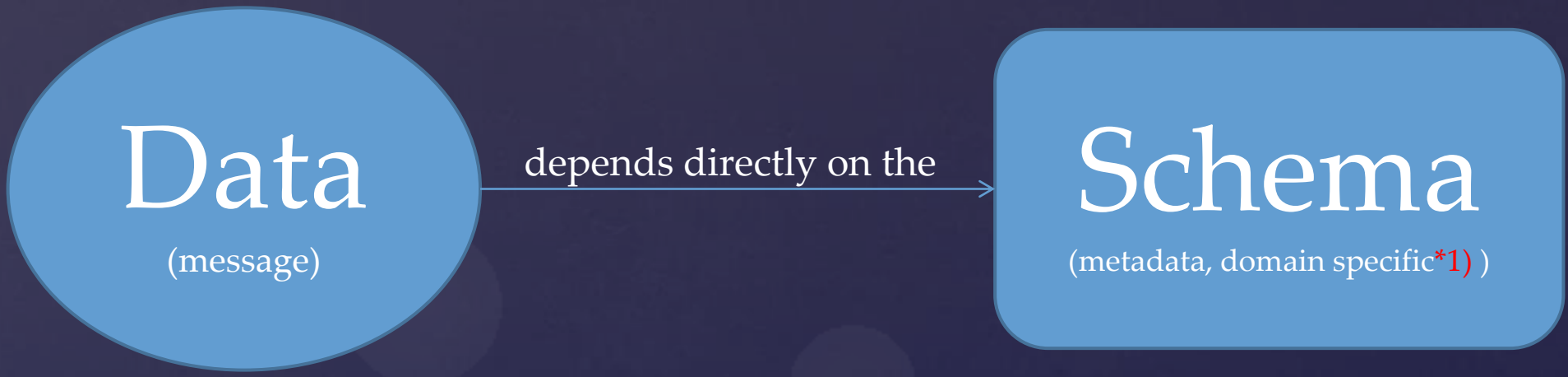
Note: XML has a successor called JSON, which is more lightweight, less verbose and web friendly

Schema or artefact based

An information standard can be ***schema*** or ***artefact*** based.

Schema or ***artefact*** refers to the type of information standard. **Or**: it refers to the style or approach taken when the information standard was developed.

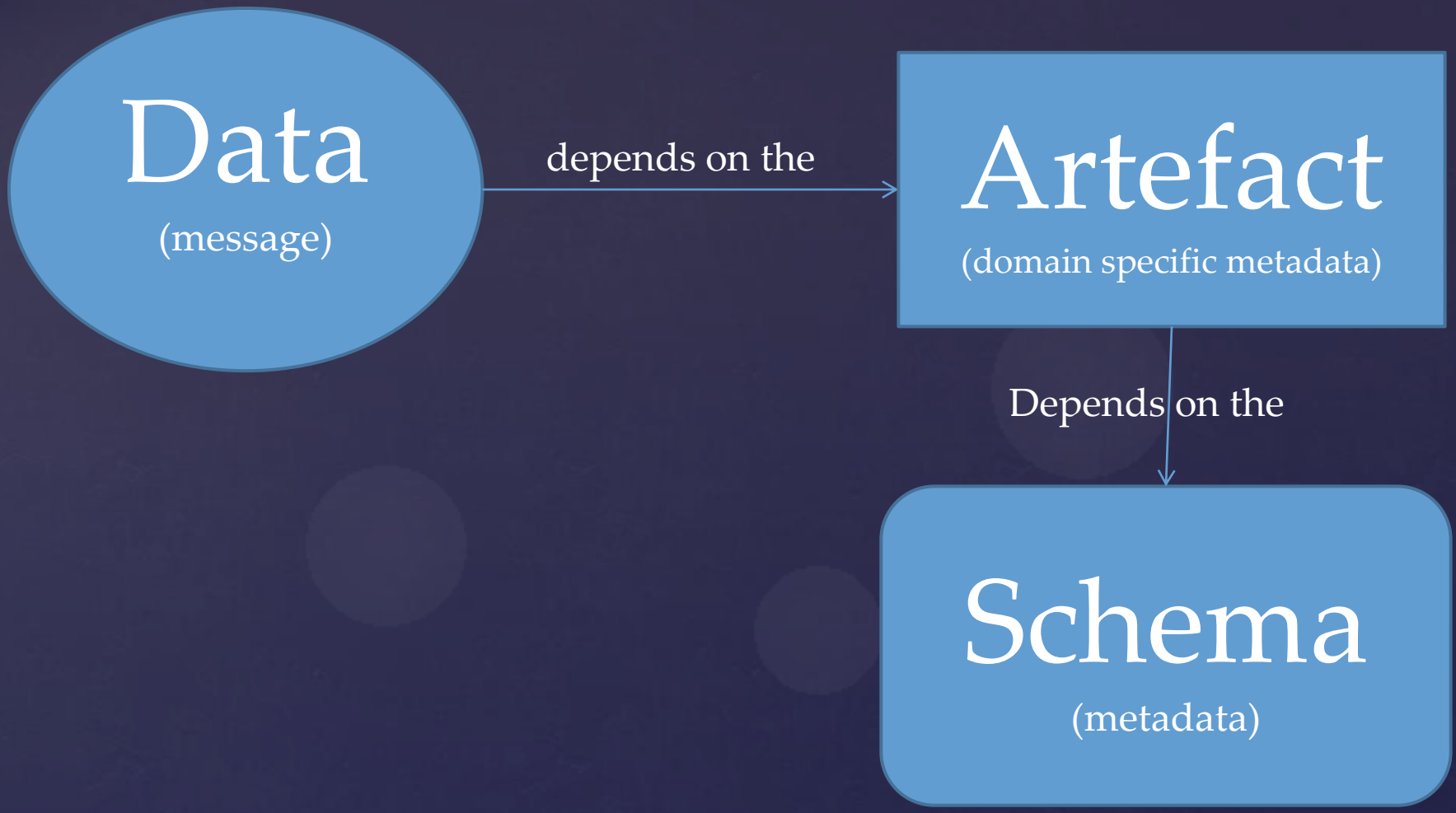
Schema or artefact based



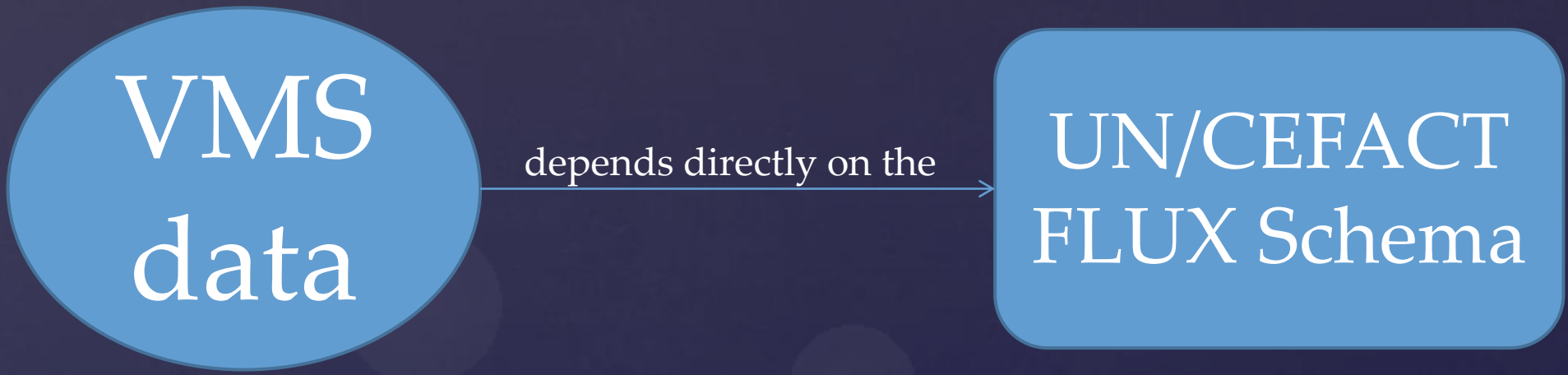
Schema based

*1) In this presentation we boldly consider all schema based standards as domain specific.

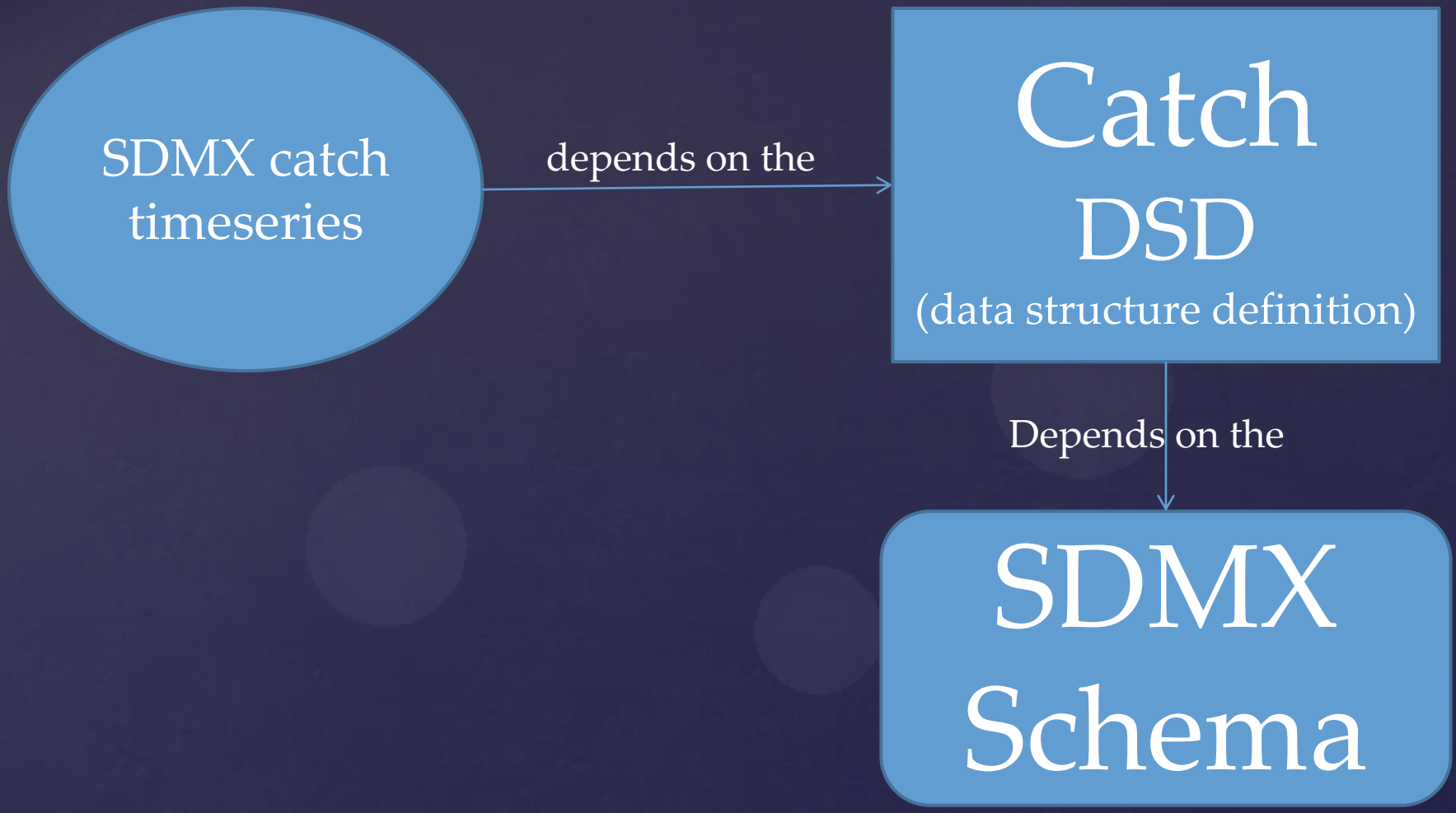
In reality this is not always the case, because a domain generic schema based standard can have domain specific extensions.



Artefact based



Example Schema based



Example Artefact based

	characteristics	pro	con	examples
schema based	often domain specific 2 layers Single Maintenance Model	less freedom pragmatic easier to implement central led harmonization	every domain has its own standard	FLUX FishFrame DublinCore FIMES
artefact based	not domain specific 3 layers Delegated Maintenance Model	more freedom 'organic' harmonization	higher complexity	DDI SDMX

	How is a codelist represented?
schema based	In the schema, often as a enumeration. This is however not practical for long lists. When it comes to long lists, often a pointer is used to the list itself, which is the same as the below mentioned technique.
artefact based	As a separate independent artefact

Master Data Management (MDM)

MDM Capacity for the management of

- Reference data
- Code lists
- Classifications
- Mappings
- Harmonization and its related workflow

Business & Transport layer

The **business** layer defines the metadata of the information (schema or artefact based). “What is in the truck?“

The **transport** layer defines how the data and metadata will be transmitted. “What are the truck and the road?”

Note: terminology copied from FLUX

implementation friendly
developer friendly
user friendly

Requirements

(related to information standards)

every standard is welcome
avoid overlap - objective
neutral - evaluate - advice
stimulate adoption
normative

(I) CWP & Information Standards

Prio1: CWP investigates, evaluates and advises on models (GSBPM, etc.) and standards (FLUX, SDMX, etc.)

Make sure all the CWP concepts & definitions are consistently applied

Prio2: Time & budget permitting, the CWP can go standard specific (e.g. FLUX&SDMX Catch DSD)

(II) CWP & Information Standards

Recommending participation of RFB agencies in FLUX and SDMX projects

Ensure coherence over models and standards like FLUX and SDMX.

Promoting the structures, concepts and definitions as valuable towards any data exchange and harmonization effort.

Recommending a MDM initiative for fisheries facilitate datasets harmonization

CWP consideration and discussion

thank you

FAO, Erik van Ingen

Swakopmund, Namibia, 25-27 February 2015