Publications Office of the European Union

Initiative for digital transformation in the Metadata and Reference Data Sector of the Publications Office of the European Union

Asset Publication Lifecycle Enterprise Architecture

Disclaimer

The views expressed in this report are purely those of the Author(s) and may not, in any circumstances, be interpreted as stating an official position of the European Union. The European Union does not guarantee the accuracy of the information included in this study, nor does it accept any responsibility for any use thereof. Reference herein to any specific products, specifications, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favouring by the European Union.

This report was prepared for the Publications Office of the European Union by Infeurope.

Document metadata

Reference	WP 1.1: Asset Publication Lifecycle Enterprise Architecture
Corporate Author	Publications Office of the European Union
Author	Eugeniu Costetchi
Reviewers	Denis Dechandon and Willem Van Gemert
Contractor	Infeurope S.A.
Framework contract	10688/35368
Work package	WP 1.1
Delivery date	25 September 2020
Suggested readers	sector management, business and technical staff, stakeholders, enterprise architects, software developers, project partners

Abstract

Public procurement is undergoing a digital transformation. The EU supports the rethinking of the public procurement process with digital technologies in mind. This goes beyond simply moving to electronic tools; it rethinks various pre-award and post-award phases. The aim is to make them simpler for businesses to participate in and for the public sector to manage. It also allows for the integration of data-based approaches at various stages of the procurement process.

With digital tools, public spending should become more transparent, evidence-oriented, optimised, streamlined and integrated with market conditions. This puts eProcurement at the heart of other changes introduced to public procurement in new EU directives.

Given the increasing importance of data standards for eProcurement, a number of initiatives driven by the public sector, the industry and academia have been kick started in the recent years. Some have grown organically, while others are the result of standardisation work.

In this context, the Publications Office of the European Union aims to develop an eProcurement ontology.

The objective of the eProcurement ontology is to act as this common standard on the conceptual level, based on consensus of the main stakeholders and designed to encompass the major requirements of the eProcurement process in conformance with the Directives and Regulations.

This document provides a working definition of what is the architectural stance and the design decisions that shall be adopted for the eProcurement formal ontology along with the specifications how to generate comprising components.

Contents

1 Introduction 5
  1.1 Background considerations 5
  1.2 EU trajectory towards semantic and linked data 6
  1.3 Target audience 6
  1.4 Document scope 6
2 Publication workflow digital transformation 7
  2.1 State of play 7
  2.2 Towards semantic technology workflow 8
3 Architecture building blocks 9
  3.1 Methodology 9
  3.2 ArchiMate elements 9
  3.3 Architecture views 14
4 Motivation architecture 14
  4.1 Overall motivation structure 14
  4.2 Stakeholders and their roles 15
  4.3 Drivers: primary 17
  4.4 Drivers: secondary 18
  4.5 Assessment: Responsiveness and processing speed 20
  4.6 Assessment: Data quality 20
  4.7 Assessment: Service quality 20
5 Business architecture 20
  5.1 Prototypical business structure 20
  5.2 Current asset lifecycle stages 22
  5.3 Actors and roles 23
  5.4 Current asset lifecycle 23
6 Application architecture 23
7 Technical architecture 23

8 Conclusions 23

1 Introduction

This document provides a working definition of what is the architectural stance and the design decisions that shall be adopted for the reference asset management lifecycle materialised as the publishing workflow used by the Standardisation Unit (SU) at the Publications Office of the European Union (PO).

In this document we (a) establish the baseline architecture, supported by the strategic and motivational information; (b) develop a target architecture guiding the transitional processes of implementing new technologies. This constitutes a natural evolution in response to changing mission needs defined by the SU management, and also takes into consideration the strategic directions proposed by PO and European Commission (EC) and European Parliament (EP).

1.1 Background considerations

Given the increasing importance of data standards for the European institutions, a number of initiatives driven by the public sector, the industry and academia have been kick started in the recent years. Some have grown organically, while others are the result of standardisation work. The vocabularies and the semantics that they are introducing and the technologies that they are using all differ. These differences hamper data interoperability and thus its reuse by them or by the wider public. This creates the need for a common data standard for publishing public reference data and models, hence allowing data from different sources to be easily accessed and linked, and consequently reused.

PSI directive [2] across the EU is calling for open, unobstructed access to public data in order to improve transparency and to boost innovation via the reuse of public data. The reference data maintained and published by the PO have been identified as data with a high-reuse potential [1]. Therefore, making this data available in machine-readable formats, following the data as a service paradigm, is required in order to maximise its reuse.

In this context the Publications Office of the European Union maintains and publishes an ever growing number of reference data assets vital in the context of inter-institutional information exchange. With regards to reference data, PO provides an ever growing number of services to the main institutional stakeholders and with the aim to extend them to a broader public; enabling active or passive participation in the reference data life cycle, standardisation and harmonisation.

1.2 EU trajectory towards semantic and linked data

European institutions set sail to adopt Semantic Web and Linked Data technologies as part of the vision to become data centric e-government bodies [cite find directives].

Many of the legacy systems in the institutions use XML data format for exchange and document formats governed by the XSD schemas. The aim is to evolve so that the existing and the new systems are capable to operate with semantic data representations using RDF [cite], OWL [cite], SHACL [cite] and other representations and serialised at least in RDF/XML [cite], Turtle [cite] and JSON-LD [cite] formats.

For this reason, the PO is already publishing the data in RDF format for over a decade using Cellar repository [cite]. And the SU, in particular, is committed to publish and disseminate the reference data in semantic formats. Next we describe the state of affairs of the SU to describe the context of the current work.

1.3 Target audience

The target audience for this document comprises the following groups of stakeholders:

- Management of the SU
- Enterprise architects and data governance specialists
- Documentalists involved in the reference data lifecycle
- Technical staff in charge of operating the workflow components
- Developers in charge of the workflow implementation

1.4 Document scope

This document aims to support SU in the transition towards the semantic technologies focusing on the architecture of the publishing workflow. The central use case is to support the asset management lifecycle presented in Section (below). It includes managing the incoming requests, editing the reference assets in VocBench3 system, then exporting the RDF data and passing them as input to a set of processes that validate, assess, transform, package and finally publish the assets in Cellar, the main dissemination platform.

This document will provide a motivational, business and application account of the workflow. Each of these accounts is limited strictly to the success scenario of the above mentioned use case and does not include possible extensions and variations which may be.

2 Publication workflow digital transformation

2.1 State of play

The SU publishes reference data in several formats, most important being XML, XSD and RDF/XML. On the technological side, the SU currently employs a legacy custom-built system for controlling and executing, in part, the asset management lifecycle operations (legacy workflow system). The system was developed using a mixture of XSLT technology, Perl and Bash scripting languages. The system was developed to execute a wide variety of conversions and transformations based on XML source files into various other formats including human readable documents.

The source data representation (XML in this case) has the primary role to serve as the only source of truth, and additionally, maintaining non-redundancy and rich expressivity. All other data forms and representations are secondary and are generated by transformation and conversion processes from the source representation.

One peculiarity of the legacy system setup is that the editing of the asset content is performed using Microsoft Excel. This is done by transformation of the content from XML representation into Excel stylesheets, which are edited by the SU documentalists, and then the stylesheets are converted back into XML form. This way a circular transformation is achieved which also serves as an integrity checking and validation mechanism. In addition XSD schema definitions are used to validate the XML source representation.

The legacy workflow system uses the file system for data persistence. In addition, this functionality is aided by a version controlling system, SVN [cite], to trace the temporal evolution of data.

Some steps in the legacy workflow are automated. The automation is based on cron tasks and SVN hooks that, upon changes in the source XML or Excel files trigger a set of conversion mechanisms. Some other steps require manual triggering and eventually parameterization intervention. The execution of the automated steps often requires assistance of technical staff or IT skills above average which represents an impediment for the non technical documentalists and a hindrance for the IT staff.

Moreover, the maintenance of this system is burdened by a technical debt that accumulated over time, because the system evolved organically based on ever flowing requests.

2.2 Towards semantic technology workflow

SU's mission regarding the technological evolution is to migrate towards Semantic Web and Linked Data technologies and representations. The maintenance of the reference data is currently done based on XML source representation and the desired transition is towards RDF based representation. For that purpose Excel and XML sources are no longer suitable and a dedicated editor is necessary.

To solve this issue, SU took the development flagship of the VocBench3[cite] system - a web-based, multilingual, vocabulary editing tool based on the SKOS [cite] model, which is built on top of RDF/S standard. Later, VocBench3 was developed to support authoring of RDFS and OWL vocabularies.

Switching to RDF-based sources and adoption of VocBench3 system implies a technological and business process disruption. The main reason being the legacy workflow system, which operates with XML based sources only and does not support RDF sources. RDF representation being only a by-product derived from XML.

VocBench3 naturally adopted a persistence based on triple stores, which are NoSQL database systems implementing the directed graph data model instead of the hierarchical or relational model of data. The relational data model is mentioned here because the Excel worksheets are based on the tabular data organisation; the hierarchical data model is mentioned because the XML is fundamentally a hierarchically organised data structure; and each of them is only partially compatible with the graph paradigm present in semantic data models.

Migration towards a new workflow that integrated VocBench3 thus requires reconciliation between file-system and database approaches to persistence. Also a paradigmatic transition to graph based data representation, from the hierarchical models of source representation, and tabular models used for source authoring is necessary.

The legacy workflow system is also lacking in validation, structural analysis (fingerprinting) and content comparison capabilities (calculating the difference between two versions of an asset). A transition would imply development of at least these new capabilities in order to maintain the current business processes.

3 Architecture building blocks

3.1 Methodology

In this document we take an enterprise architecture perspective and aim to provide several architecture aspects which are necessary and sufficient to describe the publishing process.

In developing this architecture, we adopt parts of the TOGAF [3], which is a framework for enterprise architecture that provides an approach for designing, planning, implementing, and governing an enterprise information technology architecture.

For the architecture representation, we adopt ArchiMate language [4], which is an open and independent enterprise architecture modeling language to support the description, analysis and visualization of architecture within and across business domains in an unambiguous way.

The motivation view is developed by implementing the following steps. First, TOGAF & ArchiMate were chosen as frameworks. Then some interviews were conducted with the SU management, having as a goal to elicit who are the main stakeholders and their motivations. The interview notes have been distilled and organised in ArchiMate diagrams (using EnterpriseArchitect Tool for drawing the models). Finally this report was written presenting the resulting motivation structure.

3.2 ArchiMate elements

Table 1: Overview of the relevant motivation elements [4]

Element	**Definition**	**Notation**
Stakeholder	Represents the role of an individual, team, or organization (or classes thereof) that represents their interests in the effects of the architecture.	

Table 1 continued from previous page

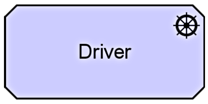
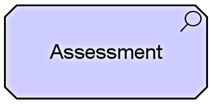
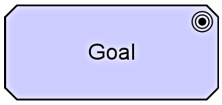
Driver	Represents an external or internal condition that motivates an organization to define its goals and implement the changes necessary to achieve them.	
Assessment	Represents the result of an analysis of the state of affairs of the enterprise with respect to some driver.	
Goal	Represents a high-level statement of intent, direction, or desired end state for an organization and its stakeholders.	

Table 2: Overview of the relevant business layer elements [4]

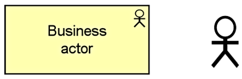
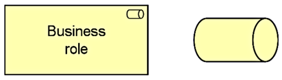
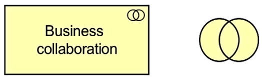
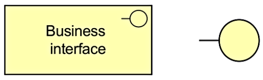
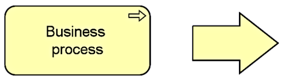
Element	**Definition**	**Notation**
Business actor	Represents a business entity that is capable of performing behavior.	
Business role	Represents the responsibility for performing specific behavior, to which an actor can be assigned, or the part an actor plays in a particular action or event.	
Business collaboration	Represents an aggregate of two or more business internal active structure elements that work together to perform collective behavior.	
Business interface	Represents a point of access where a business service is made available to the environment.	
Business process	Represents a sequence of business behaviors that achieves a specific result such as a defined set of products or business services.	

Table 2 continued from previous page

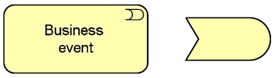
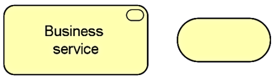
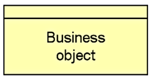
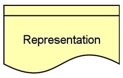
Business function	Represents a collection of business behavior based on a chosen set of criteria (typically required business resources and/or competencies), closely aligned to an organization, but not necessarily explicitly governed by the organization.	
Business event	Represents an organizational state change.	
Business service	Represents explicitly defined behavior that a business role, business actor, or business collaboration exposes to its environment.	
Business object	Represents a concept used within a particular business domain.	
Representation	Represents a perceptible form of the information carried by a business object.	

Table 3: Overview of the relevant application layer elements [4]

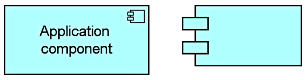
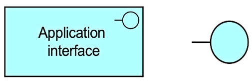
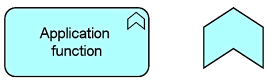
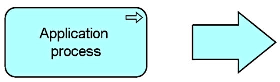
Element	**Definition**	**Notation**
Application component	Represents an encapsulation of application functionality aligned to implementation structure, which is modular and replaceable.	
Application interface	Represents a point of access where application services are made available to a user, another application component, or a node.	
Application function	Represents automated behavior that can be performed by an application component.	
Application process	Represents a sequence of application behaviors that achieves a specific result.	

Table 3 continued from previous page

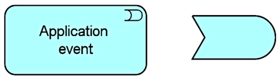
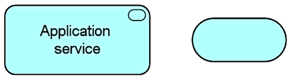
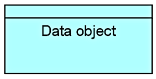
Application event	Represents an application state change.	
Application service	Represents an explicitly defined exposed application behavior.	
Data object	Represents data structured for automated processing.	

Table 4: Overview of the relevant technology layer elements [4]

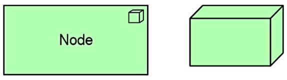
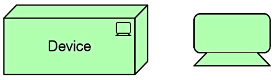
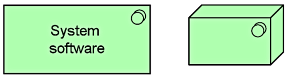
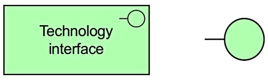
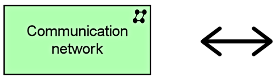
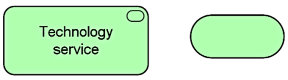
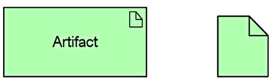
Element	**Definition**	**Notation**
Node	Represents a computational or physical resource that hosts, manipulates, or interacts with other computational or physical resources.	
Device	Represents a physical IT resource upon which system software and artifacts may be stored or deployed for execution.	
System software	Represents software that provides or contributes to an environment for storing, executing, and using software or data deployed within it.	
Technology interface	Represents a point of access where technology services offered by a node can be accessed.	
Communication network	Represents a set of structures that connects nodes for transmission, routing, and reception of data.	
Technology service	Represents an explicitly defined exposed technology behavior.	
Artifact	Represents a piece of data that is used or produced in a software development process, or by deployment and operation of an IT system.	

Table 5: Overview of the ArchiMate relationships [4]

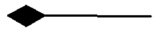
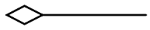
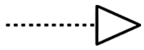
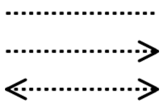
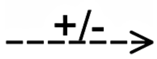
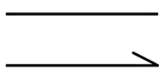
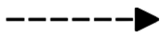
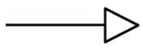
Element	**Definition**	**Notation**
	Structural Relationships	
Composition	Represents that an element consists of one or more other concepts.	
Aggregation	Represents that an element combines one or more other concepts.	
Assignment	Represents the allocation of responsibility, performance of behavior, storage, or execution.	
Realization	Represents that an entity plays a critical role in the creation, achievement, sustenance, or operation of a more abstract entity.	
	Dependency Relationships	
Serving	Represents that an element provides its functionality to another element.	
Access	Represents the ability of behavior and active structure elements to observe or act upon passive structure elements.	
Influence	Represents that an element affects the implementation or achievement of some motivation element.	
Association	Represents an unspecified relationship, or one that is not represented by another ArchiMate relationship.	
	Dynamic Relationships	
Triggering	Represents a temporal or causal relationship between elements.	
Flow	Represents transfer from one element to another.	
	Other Relationships	
Specialization	Represents that an element is a particular kind of another element.	

Table 5 continued from previous page

Junction	Used to connect relationships of the same type.	● (And) Junction ○ Or Junction

3.3 Architecture views

4 Motivation architecture

This section presents the motivation and goal structure of the Standardisation Unit. This will help determine the scope and rationale for the current architecture specification. The analysis of the motivation structure was not conducted in depth in order to constitute a decision making tool for the management but rather was aimed at accounting for the context of the publishing workflow which is the final goal of this architecture.

Nonetheless, this motivation view helps address the questions why a demand is meaningful, model crucial drivers and root causes behind the demand, actual goals and related outcomes, as well as concrete requirements for further development. In short it answers the questions to WHOM, WHY and WHAT.

4.1 Overall motivation structure

The structure of motivations, in ArchiMate, is hierarchically organised in several layers. For simplicity, we have chosen to use the top four layers: *stakeholders, drivers, assessments and goals*; leaving out the *outcomes*, *principles* and *requirements*. Figure 1 depicts the organisation of the motivation architecture. The structure starts at the top with enumerating the stakeholders, who are individuals, teams or organisations that represent their interests in the effects of the architecture [4].

Stakeholders have associated interests, concerns or drivers, which represent internal or external conditions that motivate an organisation to define goals [4].

Assessments represent results of analysis of the state of affairs with respect to some driver. They reveal strengths and weaknesses, opportunities or threats to an area of interest [4].

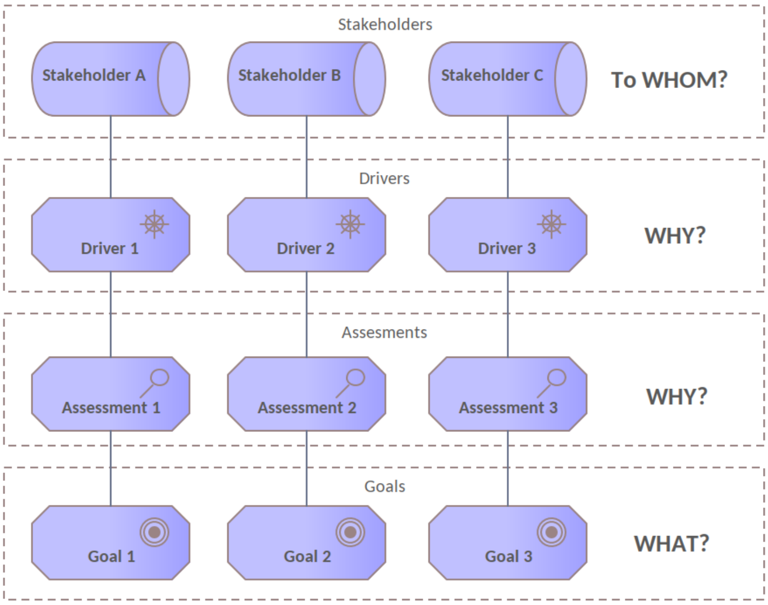


Figure 1: The layered motivation structure

Assessments are associated with goals, which represent a high level statement of intent, direction to desired end state for an organisation and its stakeholders [4].

Next we present the SU motivation structure spread over several sections.

4.2 Stakeholders and their roles

The Standardisation Unit involves multiple stakeholders. We can enumerate them but the list will be long and outside the scope of this exercise. Instead, we highlight the most important ones and in addition we group them based on the role they play in interaction with SU. In Figure 2 the roles are depicted as aggregate stakeholders in a grouping frame in the middle of the figure. Above the roles are placed the most important external stakeholders while below are enumerated the stakeholders from the PO.

The most important external stakeholders are: European Commission together with the Secretary General and all of the Commission's Directorates, the Inter-institutional Metadata and Formats committee (IMFC), the EuroVoc Committee (Group interinstitutional Lex (GIL)-subgroup EuroVoc), EU member states repre-

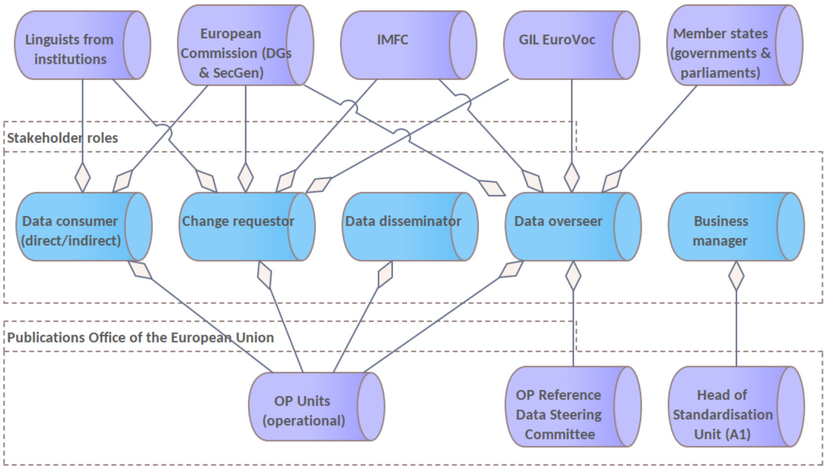


Figure 2: The layered motivation structure

sented by their governments and parliaments, and Linguists from different institutions and with a particular interest to the IATE project.

In the Figure 2 the OP stakeholders are placed in organisation bounding context, as the SU is a part of the PO and so its sibling units are not entirely external but members of the same organisation. The stakeholders within PO are the various units that use the reference data (e.g. Cellar team, OP Portal team, EurLex team, and others). A Reference Data Steering Committee is planned to be formed in the near future in order to coordinate and harmonise the published reference data. And finally the Head of Standardisation Unit who is in charge of running the enterprise.

In order to easier account for the stakeholders drivers, interests and goals, we grouped them based on their roles in interaction with SU. In Figure (above) the roles are depicted as aggregate stakeholders in a grouping frame in the middle of the figure. The roles are: data consumers, data requestors, data disseminators, data overseers and business managers. Next we describe each of the roles and then briefly enumerate the stakeholders.

The *consumers* of assets are the users that directly engage with the published assets (direct consumer) or the users of applications and services that are making use of the published assets (indirect consumer).

The *change requesters* are the agents that need and therefore requires particular content to be available as reference data.

The *data disseminators* are services and platforms where the reference assets are published for broad public consumption.

The *data overseers* are the agents that ensure that the content satisfies business needs is harmonised, coherent and complete. It is also responsible for the content correctness and harmonization among multiple stakeholders and its usefulness in broader context of application. Usually the role of data overseers is played by the standardization committees, steering committees and data stewards at large.

4.3 Drivers: primary

We have identified four *primary drivers*, three *secondary drivers* and two *internal efficiency drivers*. The distinction between primary and secondary drivers is based on whether the driver is shared between the external and internal stakeholders (in this case the business management). Figure 3 depicts the main stakeholders and their concerns, where the business manager, in this case head of the SU, has the same primary concerns as the main stakeholder roles.

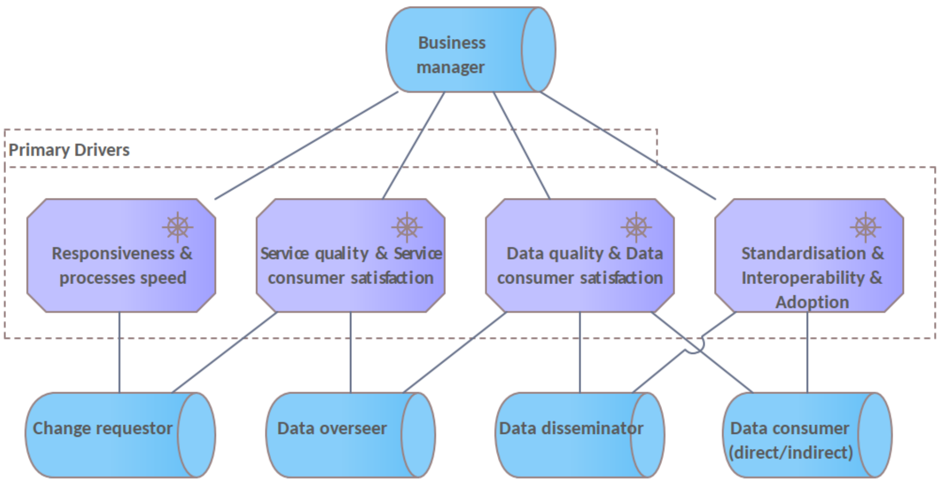


Figure 3: Primary drivers, motivating both, the internal and external stakeholders

For change requesters, the interaction *responsiveness and the speed of the asset life-cycle process* is of primary concern. The sooner the requests are processed and analysed the sooner they can be implemented, processed and published. The goal of the SU is to reach the swiftness of publishing overnight change requests, as compared to the current situation when four major publications are scheduled per year allowing also a few urgent ones in between.

The *quality of service* provided by the SU at large and the service consumer satisfaction is a direct concern for the change requesters and data overseers as primary users various SU services.

The *quality of data* is of special interest for the data overseers as they are directly responsible for this aspect and implicitly of the data consumer satisfaction. The data quality here has a wide meaning covering aspects of formal, semantic and conceptual correctness while also being timely and up to date with the business. Besides the data overseers, the data users are also interested in high quality reference data. The data disseminators are indirectly affected by the quality of the data they distribute and share this interest to a lesser degree.

The last of the primary drivers is the *standardisation, interoperability and adoption*, which is a major concern for the data disseminators and the data users. This driver covers the adoption of widely used meta-models, formally well defined models representing shared conceptualisation of major bodies and organisations, usage and implementation of national and international standards proposed by the standardisation bodies (e.g. ISO, W3C, OMG). These standards refer not only to aspects of data representation, but also to protocols, exchange schemas, validation mechanisms and other tools facilitating systemic interoperability.

4.4 Drivers: secondary

The secondary drivers are those that are important to either external stakeholders ro the internal ones alone.

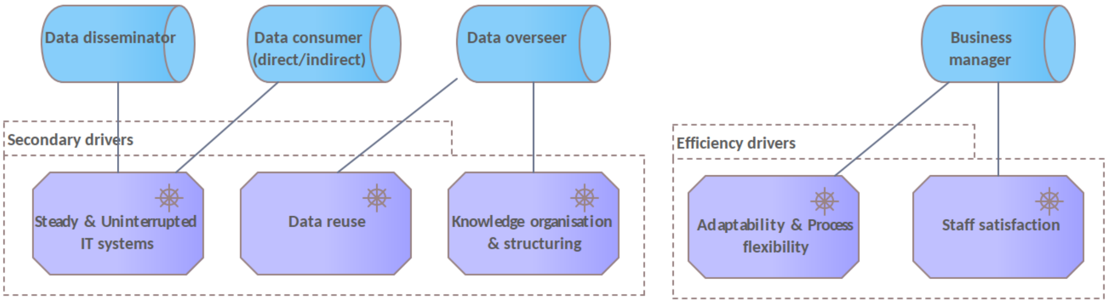


Figure 4: Secondary drivers, motivating either internal or external stakeholders

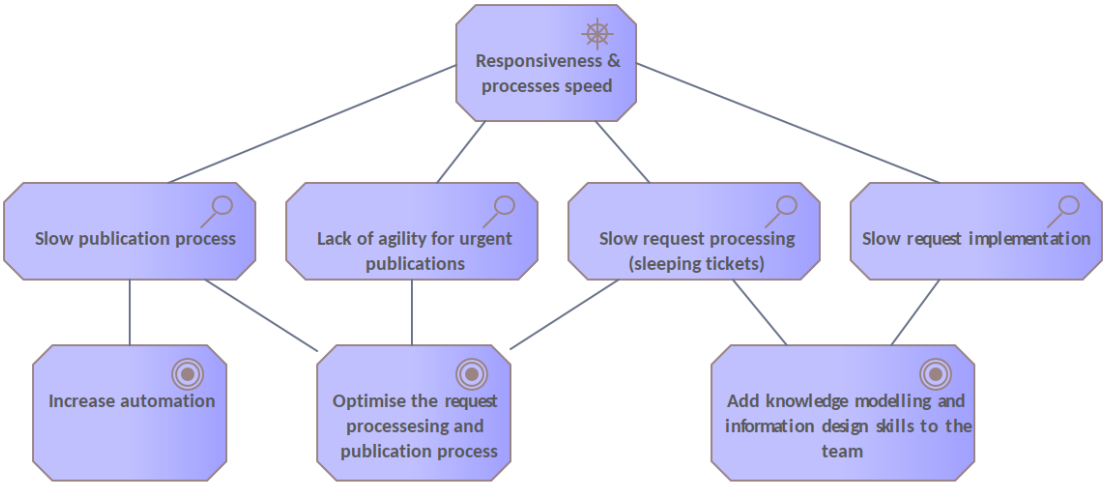


Figure 5: The assessment of the responsiveness and processing speed driver

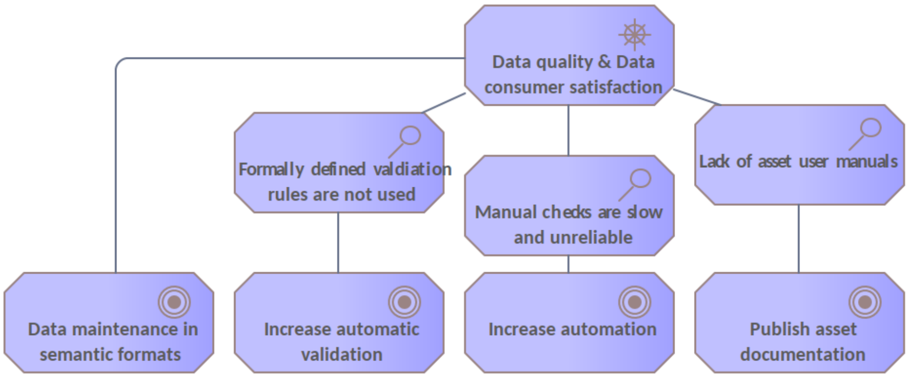


Figure 6: The assessment of data quality and data consumer satisfaction

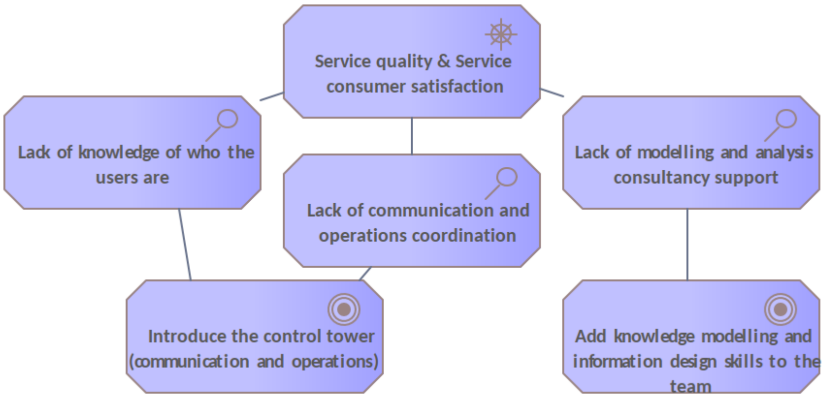


Figure 7: The assessment of service quality and service consumer satisfaction

4.5 Assessment: Responsiveness and processing speed

4.6 Assessment: Data quality

4.7 Assessment: Service quality

5 Business architecture

5.1 Prototypical business structure

Following the metaphor of layers presented in the motivation view, we decided to explain the organisation of business structure in terms of layers as well. Figure 8 depicts three layers with the most important elements of the business structure.

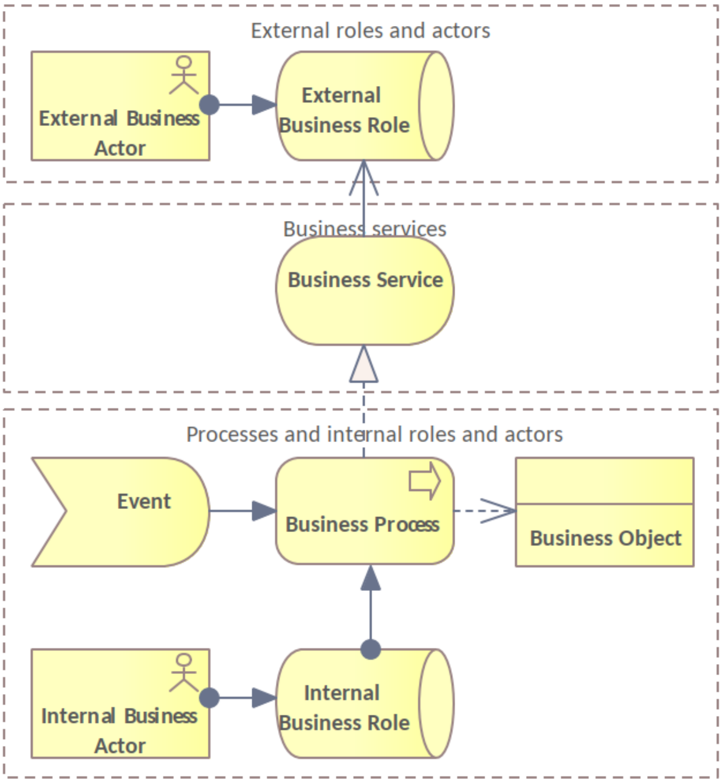


Figure 8: The prototypical business structure view

The topmost layer accounts for the external players or *actors*, which represent a business entity that is capable of performing behaviour, and *roles*, which represent skills and responsibilities for performing specific behaviour, and to which an actor can be assigned [4].

The middle layer represents the *services* that are offered by the organisation to the external players. A business service represents explicitly defined behaviour that a business role, business actor, or business collaboration exposes to its environment [4].

The lower layers accounts for the internal organisation in terms of *events*, *roles*, *processes* and *objects*. The business process represents a sequence of business behaviours that achieves a specific result such as a defined set of products or business services.

The business event represents an organizational state change; while a business object represents a (passive) concept used within a particular business domain.

In the current section we focus almost entirely on the bottom layer describing internal processes, events and roles answering the questions who shall do what and when.

Moreover, from here onwards, we start laying out two perspectives. First a baseline representing the current setup and, second, how the new processes will look like in the light of digital transformations moving towards goals identified in the motivation structure (Section 4).

5.2 Current asset lifecycle stages

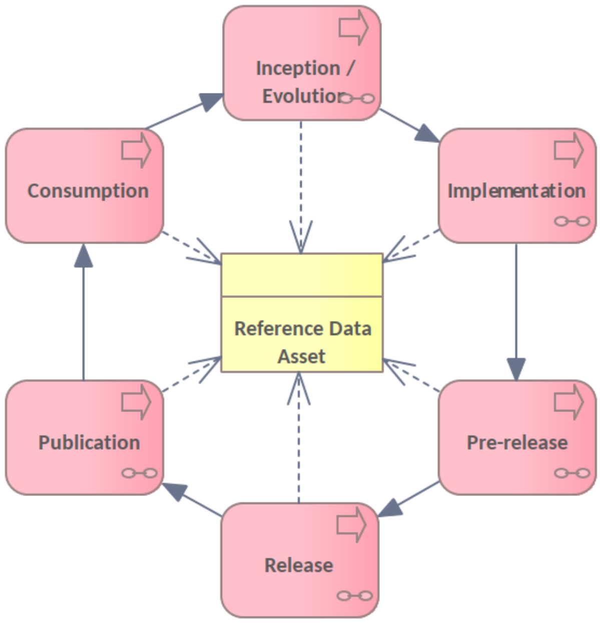


Figure 9: The current asset lifecycle stages

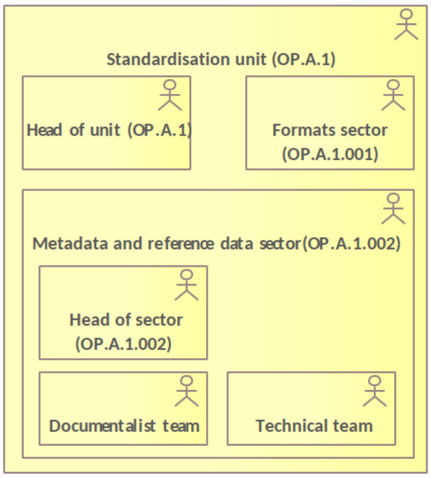


Figure 10: The actors in metadata and reference data sector

5.3 Actors and roles

5.4 Current asset lifecycle

6 Application architecture

7 Technical architecture

8 Conclusions

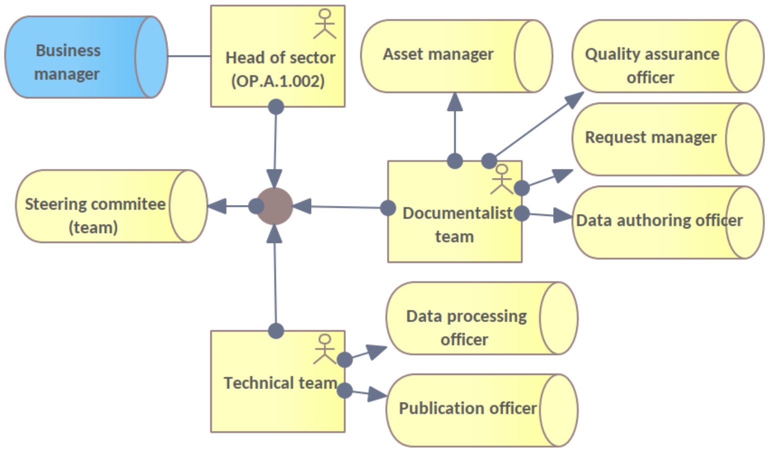


Figure 11: The internal roles in metadata and reference data sector

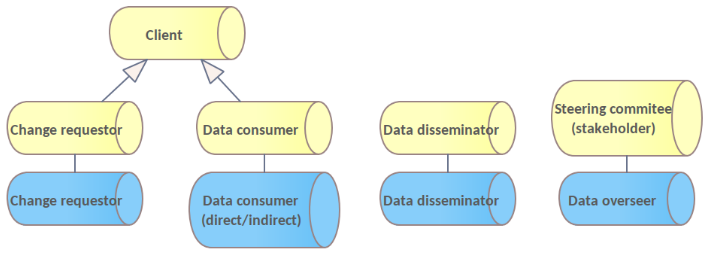


Figure 12: The external roles to standardisation unit

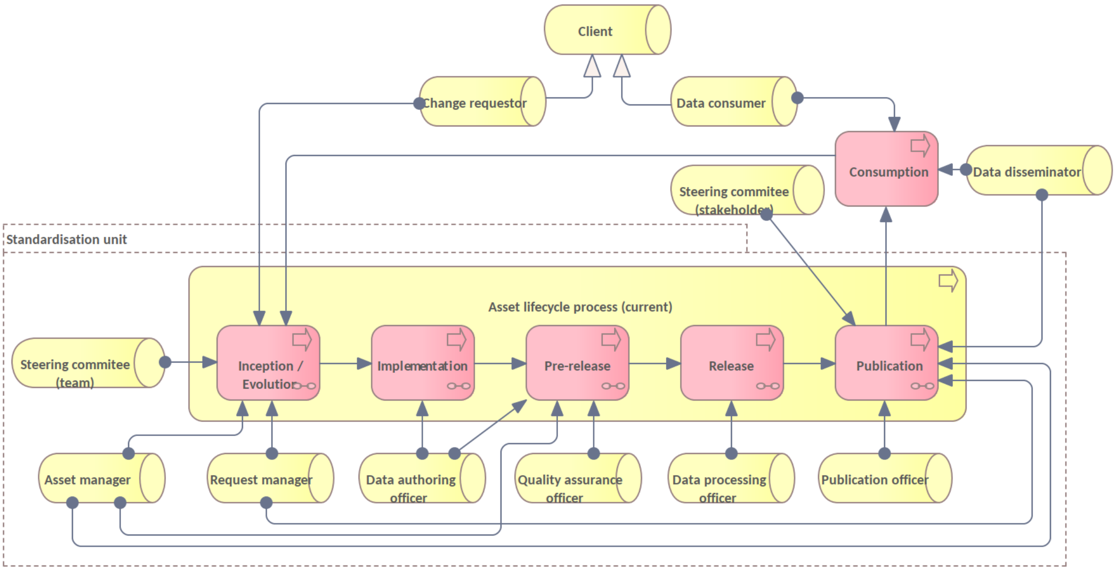


Figure 13: The current asset lifecycle stages and roles

Bibliography

[1] P. Archer, L. Bargiotti, M. D. Keyzer, S. Goedertier, N. Loutas, and F. V. Geel. Report on high-value datasets from eu institutions. Deliverable SC17DI06692, European Commission, 2014.

[2] European Parliament and the Council. Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information Text with EEA relevance. *OJ*, L 175: 1–8, 2013. CELEX:32013L0037.

[3] The Open Group. *The TOGAF Standard, Version 9.2*. The Open Group, 2018.

[4] The Open Group. *ArchiMate 3.1 Specification*. The Open Group, 2019.