# Introduction

## Motive

Recently Statistics Netherlands has defined a future vision [1.] which encompasses the concept of a Data Lake as a central capability to support all other statistical processes.

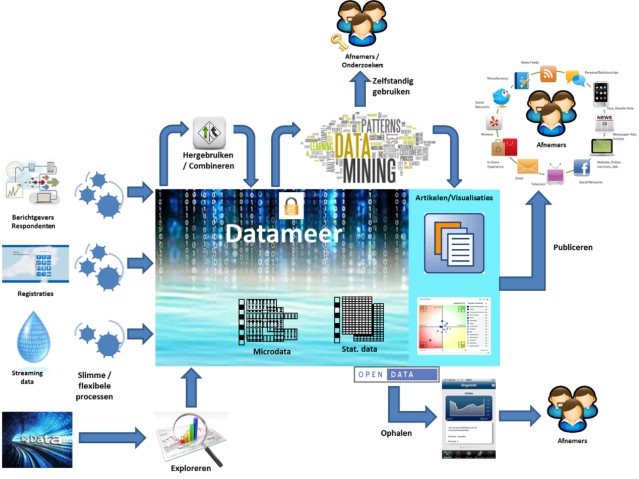
[](http://cbsintranet/werkruimten/CBS%20Enterprise%20Architectuur/Intro%20voor%20Enterprise%20Architectuur/Waardeketenplaat.png)

Figure 1

In this document we will elaborate this concept to a capability framework. Therefore, we will first introduce a general capability framework for Enterprise Information Management in Chapter 2. This framework is based on the capability framework defined in [2.]. In chapter 3 we will discuss the common patterns of use cases in this framework. Next, we will scope the Data Lake within the framework in chapter 4. Chapter 5 defines data principles which will be used as a foundation for the governance.

## Context definition

Before diving into the capability framework itself we need to draw the context of the Data Lake. The context defines the boundaries of the Data Lake with its environment and the entities which interact with it. At first we look at the information from the viewpoint of the organisation as a whole. Then it is obvious that information which is created in corporate processes are not part of the Data Lake. Because we are interested in data sources which are useful for statistical purposes only, there must be some kind of semantically information about these data sources. As long as this information is missing, for example it has to be discovered in a big data source yet, a data source can reside in a kind of sandbox. Because the presence of semantically information is a requirement the nature of the Data Lake becomes more that of a reservoir with order between structured data sources. In general a Data Lake is more associated with as set of (un)structured data sources. Therefore, from now on, we will talk about a Data Reservoir instead of a Data Lake. The next figure shows the boundaries of the Data Reservoir as defined above.



Figure 2

Now we have restricted the context to semantically structured data sources within statistical processes, in the next step we will restrict the context further from the viewpoint of the statistical processes themselves. Along with the semantically information, quality information is needed to use a data source in the statistical process. This information will be available after the data collection phase at latest. The final results of the data collection phase form the so called first steady state in the statistical information architecture. They can be incorporated in the Data Reservoir. At the other end of the spectrum the dissemination processes don’t alter the information anymore. In this processes the formats and (visual) presentation of the data sources are changed only. Dissemination processes extract the data sources from the Data Reservoir. The resulting formats after transformation within the dissemination processes are not part of the Data Reservoir. This leads to the following context from the perspective of the statistical processes.

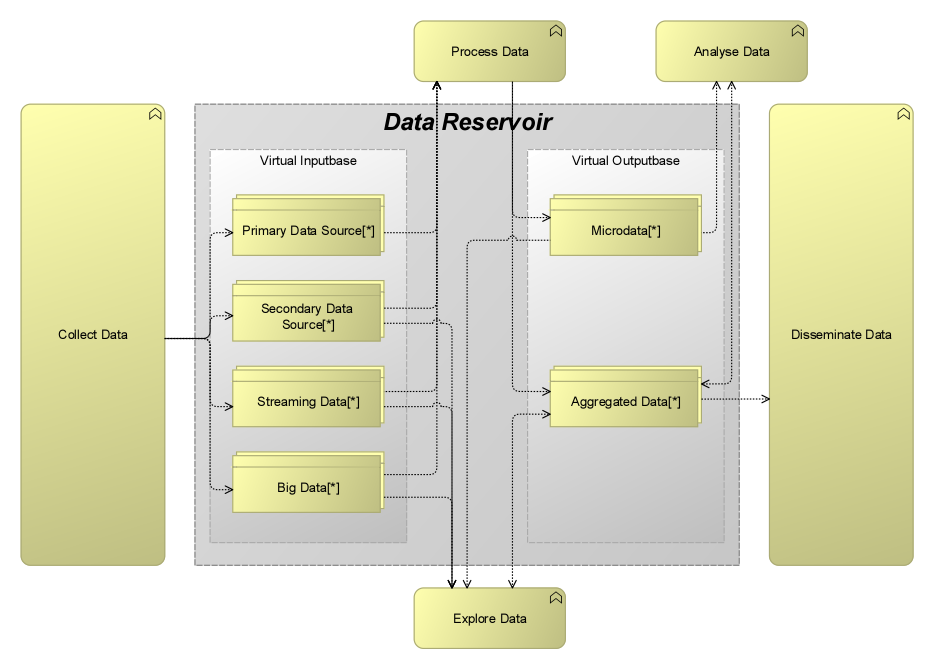


Figure 3

# Information Capability Framework

A capability defines the capacity, materials and expertise an organization needs in order to perform its core functions. Capabilities can also be broken down into more granular levels.

In an information capability framework we can distinguish 8 capabilities:

1. Data Operations: the functional operations on data sources from the perspective of a user or machine, like analysing, mining or preparing.
2. Data Usage: how the data is used to support the Data Operations
3. Manage Corporate Metadata: manage the metadata of the data sources at an organisational level.
4. Describe Data Sources: describe all aspect of one data source.
5. Organise Data Sources: organise all the data sources to make them accessible.
6. Integrate Data Sources: integrate different data sources to create new information.
7. Share Data Sources: share all data sources with data consumers.
8. Govern: govern the use of all the data sources

A detailed description of the capabilities can be found in Appendix I

If we look at the more granular level of this capabilities we can divide the capabilities in 5 vertical layers, where every layer support capabilities in the layer above:

1. Data Operations
2. Data Usage
3. Provision: how Data Sources are made accessible for the different use cases
4. Enrich: add value to the Data Sources to support the data uses cases
5. Operate: perform “physical” functions on datasets

This leads to the following overall picture of the framework:

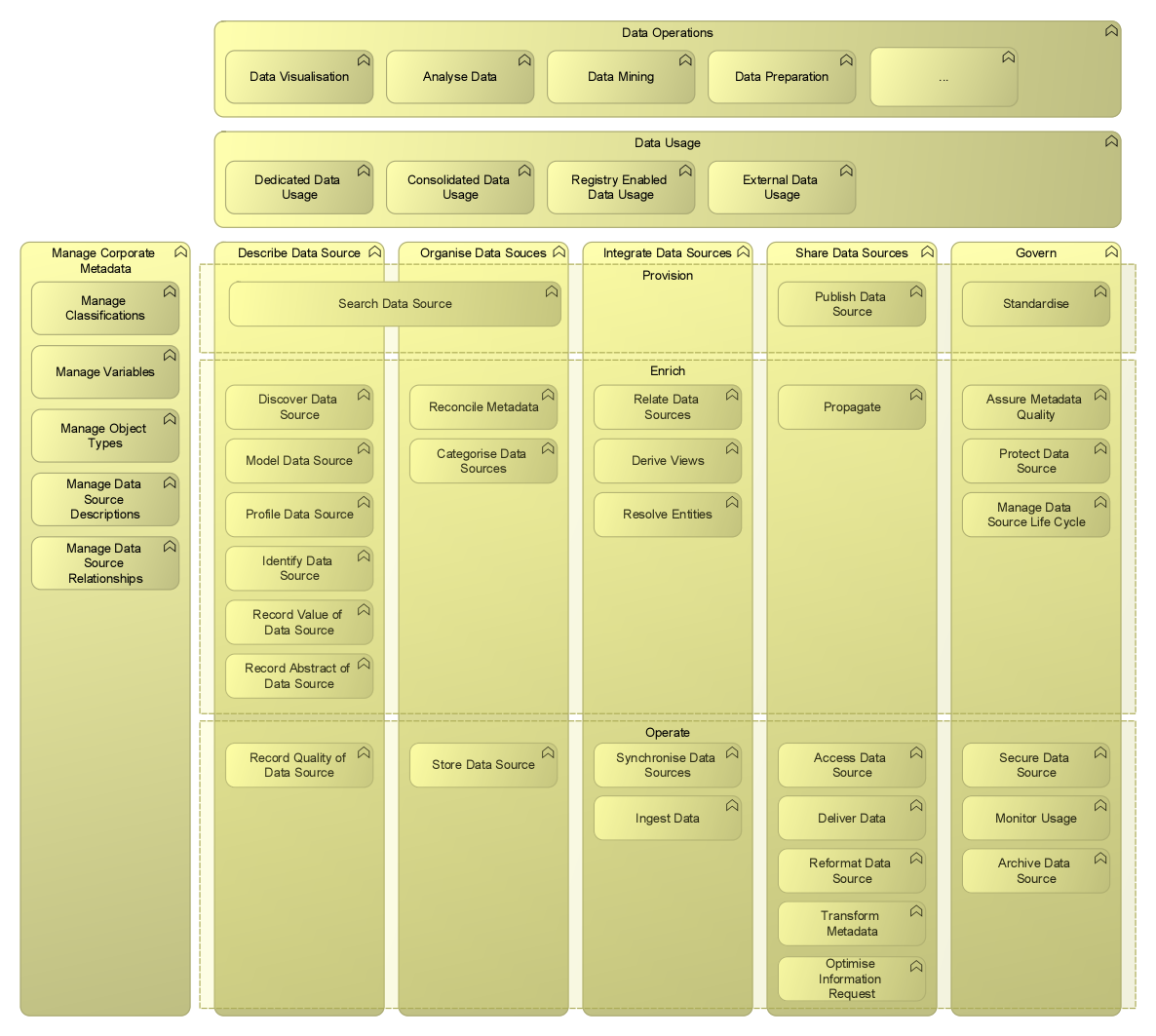


Figure 4

# Scoping the Data Reservoir

Now we have defined the context and the possible capabilities we will scope the Data Reservoir by including or excluding the capabilities which are defined in the framework. As a first step the business goals will be used to define the inclusion or exclusion criteria.

## Business goals

The vision defined in [1.] defines, amongst others, the following business goals:

1. Internal and external researchers are, without (IT-)support, able to quickly and easily search for the descriptions of all relevant information (micro data and statistical aggregates) available at Statistics Netherlands: catalogue function. Subsequently, authorised researchers have access to the data sources to analyse them and be able to produce statistics, again without (IT-)support.
2. Methods, processes and IT-systems are in place to produce statistics based on Big data sources.
3. Producing real-time statistics with pre-designed processes which are fully automated.

The results of a workshop at Statistics Netherlands about the data lake defines more goals to achieve:

1. Creating phenomenon based output
2. Faster and cheaper production of statistics
3. Improved response to ad hoc information requests
4. Improved understanding of available data sources (i.e. context information, quality)
5. Improved and cheaper access to available data sources
6. Combining data sources (cross domain)

## Data Operations

The capabilities in the data operations layer contribute to several of the goals mentioned in the previous paragraph. The majority of them are common statistical capabilities which use data sources. Accessing, sharing and combining of data sources are essential preconditions to achieve the business goals defined in the previous paragraph. These accompanying capabilities can be found in the lower layers of the framework. So far known, there is no single solution which implements all these capabilities. However, there are some statistical tools which implement, beside a set of statistical capabilities, several capabilities in the lower layers too. In theory we could select a set of tools through a best of breed approach and then integrate them into one solution. However, such a solution would result in a very complex system with tools that probably show a lot of overlap in capabilities. From this point of view we propose to leave the data operations layer itself out of scope.

## Data Usage

Keeping in mind the defined business goals, the use cases where one dedicated data source is accessed are not in the scope of the Data Reservoir. For example, to achieve the goal of combining different data sources there must be capabilities to *relate data sources* and *derive views*. This is in contradiction with the dedicated usage scenario, where there is just one data source for a specific use case.

Flexibility in combining data sources is an important requirement. Consolidated usage patterns which store data sources in one physical, often central storage facility don’t fit well to this requirement. Creating new physical data sources every time there is a change in information need is time consuming and results in a fast expanding need for physical storage in the central storage facility. We assume that users of data sources are primarily interested in information and not in the physical form or the location where data sources reside. For the time being, the consolidated approach is therefore left out of scope of the Data Reservoir.

This brings us to the conclusion that we need the registry style of use cases where different data sources can be found in a registry, that will be provided by provisioning capabilities and the capabilities to relate data sources and combine them by deriving new views. Data sources themselves reside at their physical location. In combination with the external usage scenarios, data sources which reside outside the organisation are made available too.

Because we leave the data sources outside the Data Reservoir itself (from a system perspective) the many times raised question whether processes which act on a single data source must be performed by extracting the data source out of the Data Reservoir first, can be answered. The answer is: not necessarily. After all, the Data Reservoir, is now more or less defined as a portal on top of a distributed storage layer. In that case cleansing or editing the raw inputdata could be done in place without transporting the data. Checking the process state of the source and recording the quality gives consumers enough information at the registry level to decide whether the data source suites their needs. This approach is almost inevitable when processing Big data sources.

## Final scope delineation

Because we put registry and external use in scope of the Data Reservoir, all capabilities which are used by this use case patterns (see Figure 6) are in scope too. Finally we have to decide whether the remaining capabilities *Discover*, *Profile*, *Record Value* and *Record Quality* are in scope. Although *Record Value* and *Record Quality* are not data handling capabilities it is necessary for users of data sources in the Data Reservoir to have information about the value and quality of a data sources. The value for example gives insight in the relevance of the data sources for the user. The quality gives information about the operational state of data source. With this information and accompanying business rules the statistical process can be monitored.

This brings us to the following proposed scoping of the Data Lake:

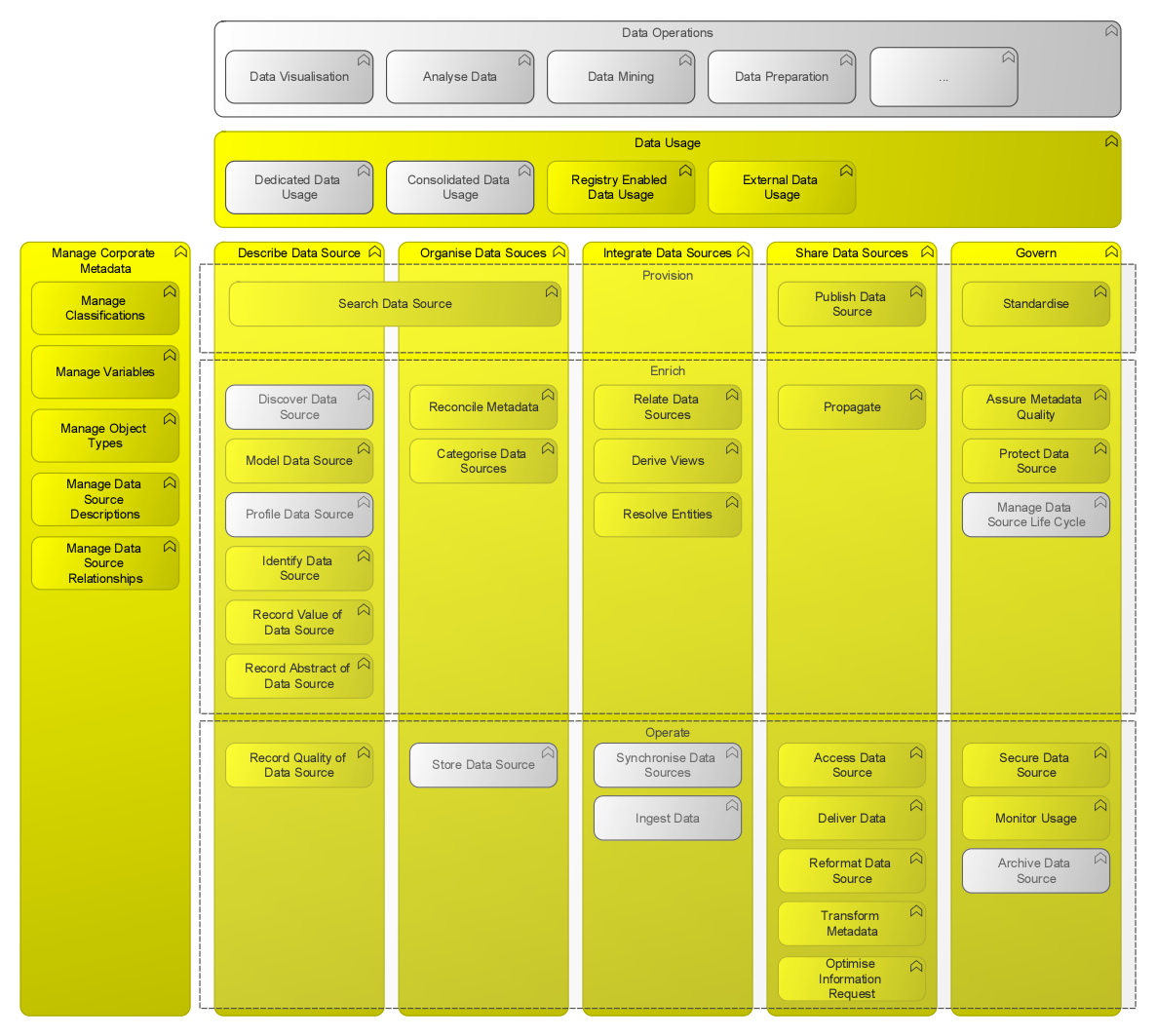


Figure 5

# Logical design

Now we have defined the capabilities of the Data Reservoir we need a logical design which groups the capabilities from a system viewpoint. This design will help to identify the boundaries of the systems which will support the implementation of the capabilities. Essentially the Data Reservoir will have a few logical layers: the provisioning layer which provides the data to the consumers, and the transformation layer which contains for example dataformat transformation capability and the capability to create new data sources from the already attached data sources. This layers are accompanied by the cross-cutting layers corporate metadata management and governance. A view of these layers is represented in the next figure. For simplicity not all proposed capabilities of the Data Reservoir are shown here.

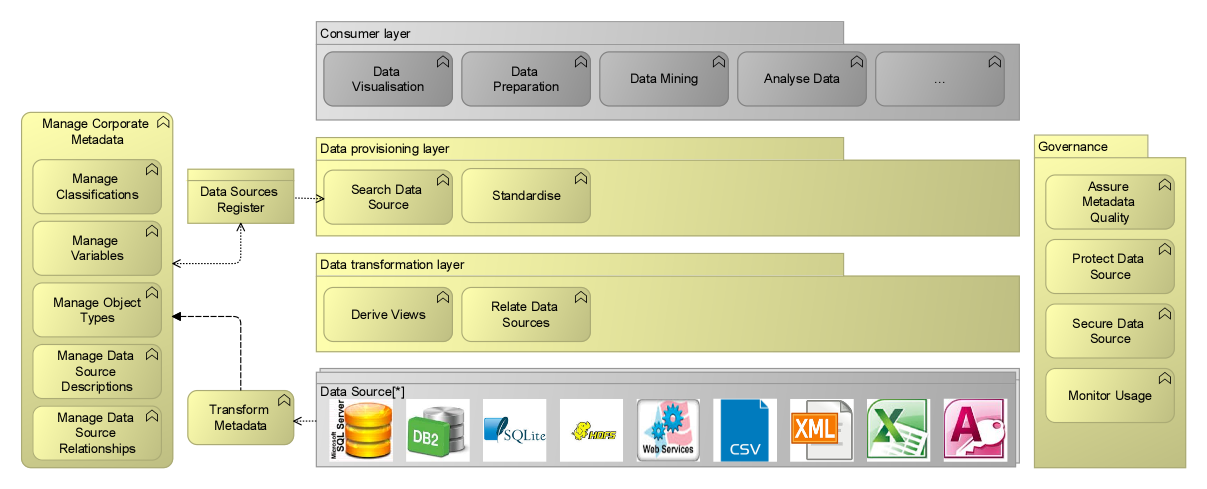


Figure 6

Because the data sources will not be part of the Data Reservoir the data sources layer is in grey. The *Reformat Data Source* capability will make it possible to use any kind of dataformat for the storage of the data source. The consumer layer will be populated with tools that actually process the data. They communicate with the provisioning layer of the Data Reservoir. This layer should provide open standards only to allow all tools which comply to open standards to actually get access to the data.

# Data Principles

To govern the data sources at an organisational level, it is necessary to define some principles. The following table shows the 5 basic principles and their consequences:

| ID | Statement | Consequences |
| --- | --- | --- |
| DAT01 | Any shared statistical data source must have an registered owner | * The ownership of data source must be recorded unambiguously * A publicly accessible overview of the ownership of data sources * The data owner is responsible for the quality of the data source * The (management) level at which ownership is defined needs to be determined[[1]](#footnote-2) |
| DAT02 | All available statistical data sources are shared on a corporate level | * Every data source needs a unique identifier * Policies and regulations regarding the handling of statistical data sources (for example, retention-time) is decided on corporate level * A publicly accessible and via the descriptive metadata searchable catalogue of available data sources * Standardisation of the descriptive metadata of statistical data sources * Standardisation of exchange formats for statistical data sources * Metadata which comes available as result of the exploration of a data source (for example a Big data source) should be recorded |
| DAT03 | The quality of a statistical data source is recorded explicitly | * Standardisation of the description of the quality of a data sources in a formal model * The origin of the data source and the process it went through must be recorded |
| DAT04 | Statistical data sources, classified as (strictly) confidential, are accessible to authorised internal and external users only | * Any statistical data source should be classified as public or (strictly) confidential in accordance with the classification employed in the security policy * The data owner is responsible for determining the authorised users * The actual authorisation to access a particular data source must be obtained via the corresponding data- owner, by means of a simple application process * IT-capabilities which store sensitive statistical data must have an authorisation mechanism * The metadata of the available data sources is accessible because of the necessity of a catalogue as defined in the consequences of DAT02 |
| DAT05 | Statistical data sources, classified as public, are accessible for all users at all times from the moment of their first publication | * Any statistical data source should be classified as public or (strictly) confidential in accordance with the classification employed in the security policy * The means of access should fit the needs of users ( omni-channel supply required ) * The substantive continuity of data sets must be ensured |

Table 1

# References

1. IV2020: Klaar voor de toekomst; Statistics Netherlands, 22-12-2014
2. Introduction to Gartner’s Information Capabilities Framework; Frank Buytendijk, Ted Friedman, Mark A. Beyer; 06-06-2014

# Appendix I



## Data Operations

The first layer of the framework consist of a diversity of potential functional capabilities. It is almost impossible to list them all, so we limit them here to some examples. In practice it will be capabilities which are used within a statistical organisation in the statistical process. All other data operations, like for example business dashboards, are out of scope. The capabilities in this layer have in common that they act on one or more data sources and produce one or more changed or new data sources.

## Usage Patterns

The next view point is the usage of data sources. What are the typical use cases ? Despite the many possible use cases and the combinations of capabilities of the framework they need , it is possible to map them to 4 usage patterns:

1. Dedicated
2. Consolidated
3. Register
4. External

Not all of these usage patterns will use all the capabilities of the framework. It is also possible to use a combination of the 4 patterns. The patterns are not non-exclusive. In the next paragraphs these usage patterns will be elaborated in detail.

### Dedicated Usage

This group of use cases usually assumes a dedicated data source with a stable data model and data storage. The data source is often used in a silod process which controls the data source itself. In this use cases there is a minimum of capabilities necessary: *model*, *access* and *secure data source*. In most cases the dedicated usage style provides capabilities optimised for the silod process. Example: a data source dedicated to one survey or a big data source

### Consolidation

This group of use cases supports the sharing of multiple data sources, which reside at multiple locations, by copying the content of the data stores to a single physical store. Often a standardised metadata model and standard structure/format is imposed. As a result it is often necessary to reformat the data source and to transform the metadata. Besides the data, a lot of metadata of the original data source has to be copied also. If it is necessary to have access to up-to-date data then this style of usage heavily relies on the synchronisation capability. With this style a repository is provided to organise the data sources and to search for available data sources. The consolidation style is strong in governance capabilities. Common technologies like data warehouses, data marts and operational data stores usually utilise this usage pattern along with ETL-tools. The within Statistics Netherlands created Data Service Center (DSC) is, to some extent, another example of an implementation of a consolidated pattern.

### Registry

Another approach to share multiple data sources, which are stored elsewhere, leaves the original data sources at their location and delivers the data on-demand. This approach is suitable in multi-channel use cases and use cases with highly flexible or ad-hoc information needs. This approach heavily relies on the organize capabilities, while they presume that the describe capabilities from the various data sources are adequate. With this style a registry is provided to organise the data sources and to search for accessible data sources. Common technologies like data federation or virtualisation utilise this usage pattern. The registry and consolidation approach are interchangeable. Which one to use depends on non-functional requirements. For example, if the frequency of access or the volume of data asked by the consumer is high, then a consolidated approach is preferred, because the accompanying tooling is able to perform at higher performance rates.

### External

In this type of use cases it is assumed that a data source enters the internal environment from somewhere else. The focus of usage is to give access to the data source where it resides. Therefore this use cases always go hand in hand with a registry style use case. Take, for instance, social media data, open data, sensor data or an external catalogue data from a supplier. The internal available metadata associated with the data source concentrates on how to share and use the data source, instead of how to change it. The external data source is only accessible as-is. In this style, both content and context remain outside of the control of the internal organisation. However, it is important to record how the data source is governed by the external owner. This use cases rely heavily on share capabilities. Access is predominately realised through webservices.

### Comparing the usage patterns

As mentioned before not all usage patterns use all capabilities. The following picture compares the 4 usage patterns concerning the capabilities they use. Each usage pattern is assigned a colour. If a pattern relies on another capability in the layers beneath, its colour appears in that capability too.

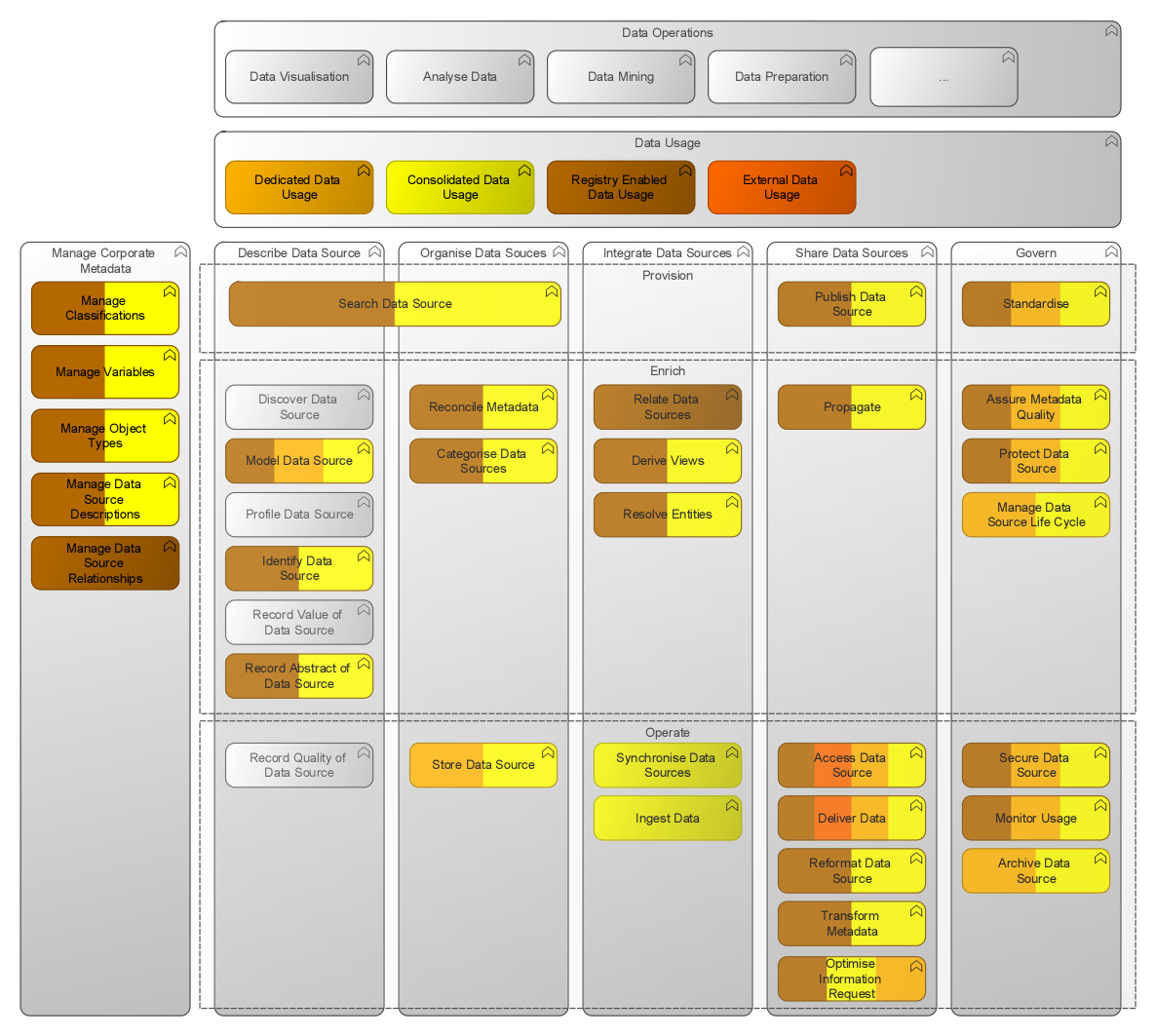


Figure 7

Note that the *Record* *Value* and *Quality* capabilities are not needed by the usage patterns. Interpreting the value and quality of data sources is not a data handling capability. Also the *Discover* and  *Profile* capabilities are not data handling capabilities. They need expert judgment or domain knowledge.

## Describe Data Source

Collect knowledge about data sources: where they are, what format they are in, what level of quality they represent and their potential value to the statistical process

### Discover Data Source

Find existing databases, content stores, applications, social networks and other external data sources that may hold information of interest to the enterprise.

### Model Data Source

Represent critical metadata describing a data source in a way that it can be understood and communicated readily. At least there should be technical metadata to use a data source in an automated environment.

### Profile Data Source

Assess the "shape" of a data source to better understand levels of quality and to further inform and enhance information models.

### Identify Data Source

Discern or establish unique identifiers for a data source that make the information findable by people, systems and other functions.

### Record Value of Data Source

Provide insight about the data source that will assist people and systems in understanding the data sources' potential value or risk in a certain context; for example, by understanding the usage of the data source in privacy oriented activities, in a specific statistical process or when combining with other data sources.

### Record Abstract of Data Source

Enable consumers of a data source to find and understand the meaning of the data source in a way that does not require them to have detailed knowledge of the underlying physical and technical implementation details.

### Record Quality of Data Source

Provide information about the current quality aspects of the data source.

## Organise Data Sources

Align and structure data sources so they can be readily found and easily consumed. They should be structured in a way that conforms to the organization's standards of syntax (format), semantics (meaning) and

terminology (use of common terms).

### Reconcile Metadata

Where appropriate, resolve and modify metadata to ensure information conforms to established terminology and classification schemes.

### Categorise Data Sources

Assign assets according to the enterprise's chosen approach to categorisation, such as taxonomies.

### Store Data Source

Physically store a data source in the selected storage location and format/structure that conforms to the agreed standards. The physical store supports readily shareable information and the non-functional equirements like for example availability.

## Integrate Data Sources

Allow independently designed data sources to be leveraged together toward a common objective.

### Relate Data Sources

Identify, via models and individual data values, links between related information assets and specific instances of data objects. In effect, this is providing the insight into understanding what can and should be integrated.

### Derive Views

Enable the joining of information from across disparate structures to produce more comprehensive views of an entity, concept or other subject of interest. Example: reusable building blocks

### Resolve Entities

In cases where consistent identifiers do not exist across multiple data sources, apply techniques for inferring relationships and remediating conflicts in descriptive attributes of entities of interest.

### Synchronise Data Sources

Support data consistency requirements by physically moving and reconciling data values across disparate systems and structures.

### Ingest Data

Acquire external data and bring it under control of the internal data management processes.

## Share Data Sources

Make data available to consumption points.

### Publish Data Source

Expose a data source to applications and to other information consumers that have registered interest.

### Propagate

For scenarios involving data consistency requirements (where two or more information sources need to remain consistent in their contents), provide for the identification of events where data is modified. Ensure that those changes are applied to related systems and data structures. Example: submit notifications of changes.

### Access Data Source

Deploy a set of protocols to connect and disconnect from the data store. Examples: ODBC, OData, etc.

### Deliver Data

Transfer information content in the designated form to the recipient.

### Reformat Data Source

Map one format/structure of a data source to another for the purpose of delivery, access or persistence. Example: from XML to csv

### Transform Metadata

Alter the semantic form of an existing information asset. Example: from DDI to SDMX.

### Optimise Information Request

Dynamically choose how to service any particular information request. Examples: optimising algorithms, caching, etc.

## Govern

Provide for control, levels of consistency, protection, quality assurance, risk assessment and compliance.

### Standardise

Agree on standards concerning metadata, quality of data sources, data source formats, governance processes, etc.

### Assure Metadata Quality

Assess the data sources with respect to the standards (metadata format, semantics) as defined. Identify and justify deviations from each standard.

### Protect Data Source

Encrypt, mask, anonimise or redact data sources deemed sensitive by security and privacy policies and controls.

### Manage Data Source Life Cycle

Manage the release of data and metadata changes. Implement rules for controlling the retainment of data sources. Example: raw input data sources should be deleted after 2,5 years.

### Secure Data Source

Assert access control to data sources, verifying that the requesting application or person is authorised to have access.

### Monitor Usage

Monitor the usage of data sources, the objects within a data source and the attributes of an object. Monitoring contributes to compliancy, privacy policies and building a knowledge base about the usage of data sources by the consumers. For example, how often is a data sources accessed by which consumer and what is the volume of data which is asked by the consumer?

### Archive Data Source

Preserve a data source and its associated metadata until it must be disposed of according to the rules of the data source life cycle management.

## Search Data Source

The ability to determine the affinity of a data source to a target concept or concepts. The search capability crosses the describing and organising capabilities. For example: in case of a repository or registry of data sources you want to drill down in the detailed description of a data sources to judge whether a data source is appropriate.

## Manage Corporate Metadata

### Manage Classifications

Manage classifications and their associated code lists across the organisation.

### Manage Variables

Manage variables across the organisation. Variables are always related to an object type. They measure a quantity (i.c. amount, weight, monetary value, etc.) or are associated to a classification.

### Manage object types

Manage the definition of object types (i.c. persons, business units, transactions) across the organisation.

### Manage Data Source Descriptions

Manage data source descriptions across the organisation.

### Manage Data Source Relationships

Manage relationships between data source descriptions.

# Document History

| Version history | | | |
| --- | --- | --- | --- |
| Version | Date | Changes | Authors |
| 1.06p | 4-12-2015 | * Introduction split in motivation and context paragraph * Context paragraph restricts the data sources to statistical data sources and data sources with semantically information only * Corporate metadata management moved across all vertical layers in the framework * Manage object types added to corporate metadata management * Assure Quality restricted to quality assurance of metadata. Quality assurance of the data itself is out of scope. * Logical design added | Hans Wings |
| 1.0p5 | 2-10-2015 | * Managing Metadata changed to Managing Corporate Metadata * Record Process Metadata added * Consolidated usage is out of scope now * Definition of capabilities moved to Appendix I. | Hans Wings |
| 1.0p4 | 24-07-2015 | Added capabilities (proposed by Hans Kleijn):   * Monitor Usage * Archive   Incorporation of the results of a workshop about the data lake at Statistics Netherlands  Inclusion of the feedback on data lake initiatives at Statistics Netherlands by Jay Devlin  Scope chapter rewritten | Hans Wings |
| 1.0p3 | 08-06-2015 | Added:   * Context diagram * Ingestion capability | Hans Wings |
| 1.0p2 | 02-06-2015 | Data Principles added | Hans Wings |
| 1.0p1 | 01-06-2015 | Initial version introducing   * Generic Information Capabilities Framework * Common Usage Patterns * First attempt to scope the Data Lake | Hans Wings |

| Active distribution | |
| --- | --- |
| Version | Distribution |
| 1.0p5 | Ilona Armengol Thijs;Paul Grooten; Harold Kroeze; Robbert Renssen; Ger Slootbeek; Jeroen Valkenburg (Statistics Netherlands) |
| 1.0p4 | Fons de Bie; Irene Salemink; Rob Warmerdam; Fong Yee Wong; IT Architect platform (Statistics Netherlands); Ivan Salomone (National Statistics Office Malta); Mark A. Beyer (Gartner) |
| 1.0p3 | Marc Lechanteur (Statistics Netherlands); Paul Brous (Rijkswaterstaat); Paul Grooten (Valid); Hans Kleijn (Nationale Politie) |
| 1.0p2 | Matjaz Jug (Statistics Netherlands); Jay Devlin (Statistics New Zealand) |
| 1.0p1 | CIO Office Statistics Netherlands; Irene Salemink (Statistics Netherlands) |

1. At Statistics Netherlands this is the head of a business unit [↑](#footnote-ref-2)