**Author:** Robert McLellan

**Version:** 44

**Date:** 2016.11.04

**Modernisation Committee on Production and Methods**

**Next Generation Data Management - White Paper working (draft)**

0

Table of ContenTs

[1 0. Introduction 3](#_Toc256000000)

[1.1 Purpose 3](#_Toc256000001)

[1.2 Document Structure 4](#_Toc256000002)

[2 1. Business Outcomes and Drivers 5](#_Toc256000003)

[3 2. Technology Platform Opportunities 7](#_Toc256000004)

[3.1 Landscape 7](#_Toc256000005)

[3.2 Key Trends 9](#_Toc256000006)

[3.3 Technology Opportunities 11](#_Toc256000007)

[4 3. Information Management Opportunities 14](#_Toc256000008)

[4.1 Linked Data and Metadata - Semantic Web advances 14](#_Toc256000009)

[4.2 Entity Resolution for Statistical Production 17](#_Toc256000010)

[4.3 Temporal Aspects of Data Management 18](#_Toc256000011)

[5 4. Information Capabilities Framework - a reference architecture for analysis 21](#_Toc256000012)

[6 5. Vision - Statistics Netherlands (CBS) Case Study 24](#_Toc256000013)

[6.1 Context 24](#_Toc256000014)

[6.2 Vision 25](#_Toc256000015)

[6.3 Capabilities: 26](#_Toc256000016)

[6.4 Scope of research 28](#_Toc256000017)

[7 6. Vision - Statistics Canada Case Study (Metadata Management) 30](#_Toc256000018)

[7.1 Context 30](#_Toc256000019)

[7.2 Capabilities 30](#_Toc256000020)

[7.3 Architecture 31](#_Toc256000021)

[8 7. Governance and Security 34](#_Toc256000022)

[8.1 Security and Risk Management 34](#_Toc256000023)

[8.2 Governance 35](#_Toc256000024)

[9 8. Next Steps 36](#_Toc256000025)

[9.1 9. Acknowledgements 36](#_Toc256000026)

0. Introduction

This White Paper on Next Generation Data Management has been created through the activity of the same name, undertaken as part of the HLG Modernization Committee for Production and Methods 2016 program. It is a collaborative effort with participation from CBS (Statistics Netherlands), Statistics Poland, Australian Bureau of Statistics (ABS), Statistics Canada, and Statistics New Zealand.

Purpose

Statistical organizations have at their core the effective use of data and metadata from diverse sources for the purpose of ad hoc and formal creation of official statistics. The advent of Big Data has seen increased focus and demand for technical capabilities in addition to those of the traditional areas of sample survey, register-based survey, and census statistics. Increased use of administrative and alternative data sources is necessitated by issues such as respondent burden (and non-response), agility and velocity, and collection cost avoidance. Finally, organizations are expected to improve their operational efficiency in the midst of budget reviews and increased demand for products (e.g. environmental accounts).

The statistical production environment uses a variety of data and metadata capture, representation, management, and transformation platforms. Additionally, data and metadata (being corporate assets) are expected to be effectively managed, retained, archived, or disposed of throughout their lifecycle. At the same time, innovation initiatives and research require flexible sandbox capabilities for ideation, experimentation, and evaluation.

The approach used to create this report has been as follows:

* Assessment of next generation data and metadata technical capabilities in the marketplace (and open-source community)
* Identification of a common reference model (e.g. the Gartner Information Capability Framework model potentially) to establish the generic functions associated with data and metadata management
* Discussion of the vision and current activities of one of the participants (Statistics Netherlands) as a valuable case study
* Proposing a series of next steps

Document Structure

This paper is organized as follows.

We start with a summary of business and outcomes and drivers that we feel are representative of the challenges and opportunities being faced today in statistical agencies across the international community. This represents the context for our exploration.

There have been significant advances in the technology marketplace in how data and metadata are captured, managed, integrated, analyzed, and visualized. An overview of technology platform advances summarizes them in the second section.

The third section discusses important advances in the area of data and metadata integration and management. We focus on three key capabilities - linked data and metadata (semantic web) approaches to describing, searching, discovering data and metadata; entity resolution; and temporal management of data series.

A useful capability framework to support the identification and creation of capabilities, functions, and services involved in information management and use is presented in the fourth section. We share an "Information Capabilities Framework" (ICF) that is currently in use within the Statistics Canada Enterprise Architecture area, based on a model from Gartner. This reference framework can provide a useful tool understand capability mapping to underlying services and functions.

A Case Study highlighting the work under way at Statistics Netherlands is discussed as a "real world" example of defining and realizing a next generation data management vision, with a particular focus on data lakes at the core of the study.

A second Case Study shows work currently in progress at Statistics Canada in metadata management, building on semantic web and graph database technologies to deliver a new solution (Picasso).

A section on Governance and Security gives a short overview of key areas of focus.

The last section discusses proposed next steps and linkages to existing or proposed HLG work streams and areas for further work.

1. Business Outcomes and Drivers

It is important to identify the key business drivers for NSO's that necessitate new approaches - put simply, what are the business values and impacts of new approaches to data management?

In our discussions we have identified a number of strategic outcomes:

* Agility - the ability to respond to quickly to emerging stakeholder needs to ensure relevance (e.g. responding to "deliverology" requests for outcome-driven government service delivery)
* Discovery - supporting the creation of innovative new statistical products based on data exploration and experimentation
* Efficiency - effective use of storage, compute resources (IT Infrastructure)
* Quality - enhanced repeatability, reliability, reversibility, interpretability, coherence arising from effective data management
* Flexibility - support for alternate means of production, taking advantage of new and emerging data sources - Internet of Things, administrative data, satellite / imagery, social media data, web-scraping
* Scalability - ability to deal with large data sets in a variety of activities - e.g. big data analysis, record linkage, etc.

There are a number of business values to be realized from new approaches, including:

* Effective balance been complete centralization and normalization (e.g. warehouses) on the one hand and completely decentralized and locally managed approaches (on the other hand)
* Quality - the management of data lineage throughout acquisition, processing, analysis, and dissemination
* Efficiency by leveraging centralization and standardization for "staging and access", and enabling powerful consumer-driven processing and analysis -
* Lifecycle management - addressing the need to manage the lifecycle of data (including destruction, retention, and disposition)
* Flexibility - the ability to integrate, offer mix and match approaches to data exploration in an efficient way (instead of copying everything everywhere in an uncontrolled fashion)
* Innovative new products - Innovation delivering new products and means of production (e.g. ancillary data approaches, data visualization, ...)
* "Plug and play" of data and production components in conjunction with CSPA, GSIM, LIM, DDI, SDMX, others

2. Technology Platform Opportunities

As a consequence of the increasing availability of data, the technology landscape has seen significant change in the technology offerings available to support the business capabilities for storing, processing and analysing data. The advent of big data has resulted in the proliferation of technology that is able to underpin next generation data management needs.

This section will present a high level outline of the broad technology landscape, the key trends and the opportunities that are relevant for next generation data management solutions.

Landscape

The next generation of data management solutions by definition will require technology solutions beyond those offered by traditional database solutions. Business demands to drive greater value from the growing volume of data have seen the emergence of a variety of new platforms that support the storage, processing and analysis of big data. Their value proposition is in the ability to align with the innovation, speed and agility required to deliver more sophisticated applications to business at a faster pace than we have been able to in the past.

Rather than trying to compile a comprehensive list of products and services which is continually growing, we have referred to a Forrester Research paper – (The Patterns of Big Data).

Across organisations involved in next generation data management initiatives, there have emerged a number of architecture patterns that are being adopted for big data implementations:

* Data warehouse augmentation where the data warehouse remains the core component of the data architecture, but some of the secondary analytics is shifted to a big data platform.
* Data refinery where a distributed data hub is used for data staging but data is persisted and analysed within the current relational infrastructure.
* All-in-one where all data is persisted in the big data platform with applications and analytics also integrated into the same platform which becomes the single authoritative source.
* Hub-and-spoke where a physical, logical or distributed hub surfaces the data that is needed in the format that it is needed to be consumed by the application or user, which can then be analysed in the hub, but more typically moved out to another data store (spoke) for further processing and analysis.
* Standalone package that is generally a big data analytics tool that meets a specific departmental need but will have limited capabilities for supporting next generation data management needs.
* Streaming analytics where support is required to address the processing and analysis of high velocity or fast data in near/real time.

Analysis of the architecture patterns reveals a number of data management technology building blocks that exist within the patterns:

* Distributed data hub that underpins the architecture by providing a low cost persistent data layer that meets requirements for availability, security and recovery and exposes data for transformations and analytics.
* Data services that include contextual services for data quality, metadata management and modelling and delivery services such as federation/virtualisation, transformation and security services.
* Enterprise data warehouse / analytical data stores that provide relational data storage solutions for optimised SQL analytics.
* Big data analytics that provides data operations and analytical tools that interact directly with the distributed data hub.
* BI and analytics tools that are the traditional tool sets that operate on the data that has been surfaced from the data hub for processing and analysis traditional operational or structured analytical data stores.
* Data science workbench comprised of the tools used by data scientists to perform data exploration, model development and management, and data mining and advanced analytics on the data that resides in the distributed data hub.

The building blocks that will be needed can then be determined by an assessment that aligns them with the outcomes required to deliver on business strategies.

This in turn informs the appropriate technology architecture pattern to apply, which can then be used in the solution design and vendor selection process where the technology decisions will be made.

This method prescribes an approach that allows organisations to identify the building blocks that they require for their next generation data management environment and then map that to a technology architecture pattern that can be used as the basis for determining the appropriate technology choices.

The marketplace has seen a proliferation of products that have been developed to manage data in new ways – Figure 1.

To avoid the pitfall of making a technology driven decision in a very complex marketplace the building blocks approach is a viable method to help navigate through the overwhelming choice of technologies that now exist, and guide the right decisions for delivery of the organisation strategy.

Key Trends

The relational database era was driven by a model around standardisation, both in terms of a standard approach to model data (relational) and a single method (SQL) of then accessing it. Initially the relational database market consisted of many vendors producing products that would distinguish themselves from other vendors, but over a period of time the capabilities of these products started to converge and we are now left with just a few dominant players. However this might not be the pattern that repeats in the big data era, as the drivers for adoption are quite different. Specifically, the business demands to derive insight and value from the vast volumes of data being generated is now able to be met by technology solutions that are able to store and process large volumes of data using cost effective scale out architectures. Current database offerings target very specific workloads in a horses for courses approach.

In considering the key trends that might influence the next generation data management space, it is interesting to review the path that big data has taken. Initially it was the major internet companies like Google, Yahoo, Facebook etc that were faced with having to store and process unprecedented volumes of data, which led to them needing to build their own big data technology infrastructure to support their business needs. It was also a period where the open source movement was gaining some momentum, so resulted in these technologies being shared within active open source communities. This in turn led to big data start-ups adding to the technology options becoming available in this new market. At that time the early adopters of this new technology were the digital native companies that had no existing legacy infrastructure and were facing similar needs to the large internet organisations, so the big data path was an easy choice.

If we fast forward to a few years ago when big data has been around for a while, main stream adoption by a broader set of organisations across industry has been slower and trickier. The reason being that unlike the digital native companies, organisations like statistical agencies have a lot invested in existing technologies that on the whole do the job that is needed. Although there is an acceptance that modernisation of the statistical production process and overall statistical agency capabilities is necessary, there has been a cautious approach to moving to a new technology platform and certainly no desire to rip and replace critical production systems. The trend has been to adopt an evolutionary approach, that has involved starting pilot projects or in some cases implementing point solutions using some of the new technology offerings in the market, through to more enterprise focussed considerations around data lakes. These efforts have helped build internal competencies and created relationships with various vendors in this landscape.

In terms of what has been happening on the vendor side, many of the early products were positioned in the infrastructure layer in the domain of developers and engineers. Some of these products have been around for some time and have matured to be enterprise capable offerings. However there has still been significant innovation of products in the infrastructure layer, in large part due to open source activity.

Probably the most significant trend in this space has been the SQL support that big data platforms have provided, from Apache Hive and Cloudera Impala for Hadoop through to the SQL capabilities available for the various NOSQL databases. This has opened up the accessibility of data in big data platforms to a vastly wider audience that are now able to use familiar tools and a familiar language, hence increasing the return that can be gained from big data investment.

A more recent trend that has developed over the past ~18 months has been Apache Spark interest. The significance of Spark is the capability it brings for in-memory processing providing major performance improvements over Hadoop’s MapReduce. Although Spark is not the only offering in this space, with other leading products with in-memory capability like SAP Hana (among others), Spark has been the product that has achieved some traction. Following on from Spark but also related to a capability provided within Spark is the whole area of machine learning. Whereas the innovation around big data infrastructure has relevance for developers and engineers, machine learning sits in the analytics layer that is the domain of data scientists and analysts. While the algorithms being used are not necessarily new developments, the advancement of the technology has provided the opportunity for them to be able to be applied to vast amounts of data to derive insights both cheaply and quickly. Organisations ability to leverage machine learning capabilities is an important part of realising the promise of big data. The development and application of machine learning will be an important consideration for statistical agencies. As the technology improves consideration will need to be given to the automation that can be employed in data analytics and statistical production and how that will relate to the roles and responsibilities of data scientists and statistical analysts in the business process.

A more recent trend sees big data offerings appearing further up the technology stack. As some of the core big data infrastructure layer challenges are being solved and much development in the analytics layer has occurred in recent times, the big data application layer that aims at business users and consumers is the area that is seeing new products emerging. This trend provides the opportunity for organisations to leverage big data without having to necessarily deploy or even understand the underlying big data technology. However this particular trend is likely to have less relevance for statistical agencies as ours is not a vertical industry that is being directly targeted.

In summary, the trends around the technology to support the next generation data management needs have gone through a phase where there has been a lot of innovation in the development of platforms provided in the infrastructure layer focussed on storing and processing the increasing volumes, variety and velocity of data – beginning with Hadoop and the rapid expansion of open source projects in that ecosystem and the proliferation of NOSQL database offerings as well as the incorporation of big data capabilities into existing vendors products. The maturing of these infrastructure layer platforms has led to the next phase with the emphasis shifting to the analytics layer where the dominant trend has been the race to provide machine learning which will be a key next generation data management capability for allowing organisations to create more efficient business processes and find new ways to create value from the insights to be gained from their data.

Technology Opportunities

The increasing availability of data presents both challenges as well as opportunities from a next generation data management perspective. The challenge of being swamped by the growing volume of data needs to be met by the opportunity of using that data to deliver better services and insights to our customers. There will be challenges around collecting, storing and accessing data that need to be resolved, but the key challenge is in determining what data is of use to the business as the real opportunity is not just about having the infrastructure to collect and store data but to be able to contextualise it and understand it in order for the users of the data to gain value from it.

Next generation data management technology is needed to support business needs that are demanding faster access to a larger volume and greater variety of data while having the agility to respond to changing business requirements.

This new landscape has created opportunities for technology to be developed to address key data management challenges around:

* Integrating diverse and complex data
* Improving performance
* Securing private data
* Coping with high data volumes
* Handling unpredictable workloads
* Delivering high availability

Although we cannot predict the future use and growth patterns for big data, current trends would indicate that it is important for statistical agencies to understand the current opportunities by gaining an understanding of what is involved the development of a next generation data management capability. Fortunately big data is supported by a very scalable model. This provides an opportunity to make a start without having to make a large capital investment, particularly when using the flexibility of cloud services.

First steps could start with finding a practical and cost effective use for big data. If the evaluation finds insights and determines that there is value, the flexibility of the cloud model will allow for it to be easily scaled out.

For investment in next generation data management technology to be viable we must understand the sort of use cases that now exist in this new data landscape from which we can derive business value. The common themes that come through from research papers, organisation/executive surveys and case studies identify a number of use cases that organisations are leveraging their next generation data management capabilities to derive business value from data include:

* **Advanced analytics that extends beyond current BI reporting and OLAP capabilities** to include data mining, natural language processing and machine learning that will be required for data discovery and analysis in big data environments.
* **Data exploration of new data sources** to discover new business opportunities for interacting with customers and partners.
* **Applying new big data techniques to analyse data that we already have** to derive new insights that current analytical tools have been unable to reveal.
* **Start leveraging unstructured data assets** using algorithms for text analytics to gain insights into text laden business processes or to perform sentiment analysis on customer oriented interactions.
* **Extending existing customer analytics to include social media data** from both social media sites as well as existing channels that enable customers to voice opinions and facts that could identify new customer segments, proactively manage product or service issues and monitor sentiment drivers.
* **Improve existing analytics by incorporating big data sources** where the relevance for statistical production is its dependence on sample size could be enlarged and broadened while addressing the respondent burden issue.
* **Harnessing real-time data** to augment or replace existing statistical output in the areas that real-time insights are being demanded, by leveraging big data streaming solutions.

Abbildung 1 Figure x. Big Data Landscape

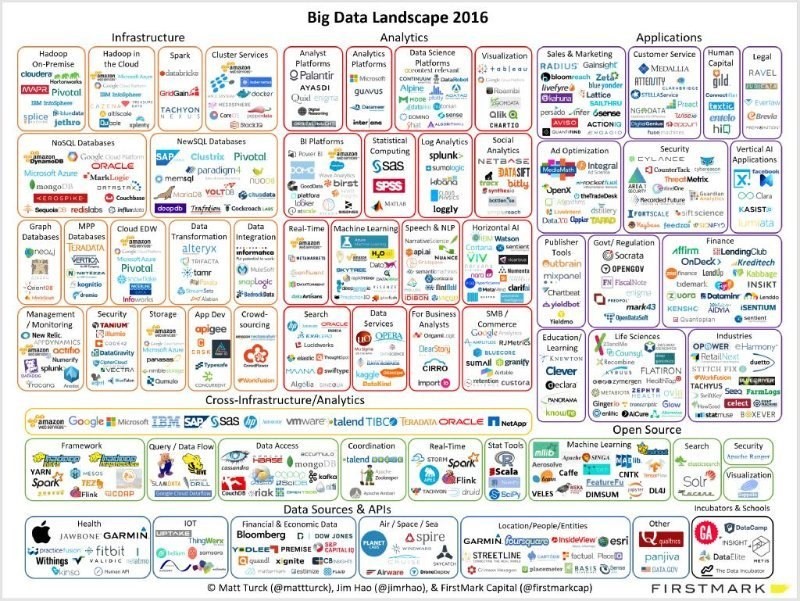


Figure x. Big Data Landscape (Source: Firstmark Capital). xxxxx)

3. Information Management Opportunities

Linked Data and Metadata - Semantic Web advances

There are five key business capabilities identified in the Information Capability Framework (below) that are key to successful data management and use - describe, organize, integrate, share, govern. Within an Agency (or between them), we wish to ensure that we can put in place solutions that deliver these capabilities, linked to powerful data exploration, analysis, and visualization.

If we look externally we see architectural patterns writ large that demonstrate collaboration through shared data and metadata, resulting from experiments, to support the creation of knowledge and its products in support of bioinformatics, genetics, infectious disease, and many other areas. There are many projects on the web, with tool, technology, and collaboration projects CEDAR (<http://metadatacenter.org>), Protegé (<http://protege.stanford.edu>), ISA tools (<http://isa-tools.org>), Cytoscape (Network Data Integration, Analysis, and Visualization <http://www.cytoscape.org>).

The same techniques being used more broadly in the Semantic Web can be effectively applied within an Agency. We provide a quick overview of semantic web opportunities on the broader web with the understanding that these are also very relevant within an Agency. There are a number of recent examples from Statistical Agency activities (e.g. INSEE presentation at Standards Workshop, September 2016).

Linked data provides a conceptualization to integrate seemingly disconnected Web sources into a coherent and accessible database sometimes referred to as "Web of Data". For data to be considered Linked Data it has to satisfy the following accessibility principles:

* Data is available on the Web with an open license
* Data is available as machine-readable structured data
* Data is available in a non-proprietary format
* Data is published using open data standards from the W3C (RDF and SPARQL) to identify things
* Data is linked to other data

The more accessibility principles any given data collection satisfy, the easier it is to understand and use.

Linked Data is based on a number of standards: Uniform Resource Identifiers (URIs) and their extension Internationalized Resource Identifiers (IRIs) for identification, the Hypertext Transfer Protocol (HTTP) for access, and the Hypertext Markup Language (HTML) for content format. Unlike the traditional Web, which relies solely on linked hypertext documents (HTML) with no explicit semantics, the links in the Semantic Web of Linked Data are more generic entities described by RDF and satisfying the following entity design principles:

* Use URIs as names for entities
* Use HTTP URIs so that people can look up those names
* When someone looks up a URI, provide useful information, using the standards (RDF\*, SPARQL)
* Include links to other URIs, so that they can discover more entities.

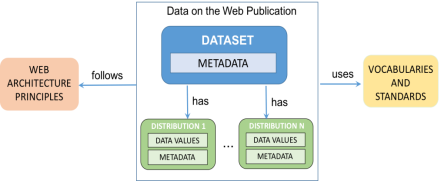
URIs provide a simple and extensible mechanism for identifying Semantic Web resources, i.e. entities. They are a generalization of the URLs used in the traditional Web to identify pages. HTTP URIs are those resolvable, i.e. localizable, on the Web. Linked Data uses only HTTP URIs, rather than their alternatives such as URNs and DOIs. In addition, the URI should resolve to some useful description of what the entity is, either human-readable (in HTML) or machine-readable (in RDF). Finally, in order to be truly Linked Data all data collections should be linked to others.

In the Web of Data, it's important to distinguish between the real-world objects and their descriptions. Since URIs can identify both, we need to have separate URIs for the entities themselves and for the metadata describing them. Being able to distinguish between the two is critical for the semantic consistency of the Web of Data. RDF provides semantically rich links for the Web of Data. HTML links connect documents, but the meaning of its links is not explicit in the model and is left to the human user to interpret. RDF in contrast, provides typed links which state explicitly the meaning of the connection.

The RDF graph data model is a collection of triples of the form subject, predicate, object. Each triple can be viewed as an assertions saying that a relationship (the predicate) holds between the subject and the object.  
RDFS is a simple ontology language, or vocabulary, built on top of the RDF data model. Similarly to RDF, an RDFS ontology consists of a collection of triples, but this time subject and object are RDFS resources. In other words, the RDFS graph is a collection of assertions between resources. In addition, properties in RDF are grouped into a class, which means they can be extended. The RDFS mechanism for extending properties is a special property called subPropertyOf.  
OWL ->  
SPARQL ->  
More recently, there has been a development to make Linked Data more usable. This resulted in the FAIR data principles, which state that:

* Data should be Findable, i.e. data should be uniquely and persistently identifiable
* Data should be Accessible, i.e. data can be always obtained by machines and humans
* Data should be Interoperable, i.e. (Meta) data should machine-actionable, their formats should utilize shared vocabularies and/or ontologies
* Data should be Re-usable, i.e. data should refer to their sources with rich enough metadata and provenance to enable proper citation.

Building on the FAIR data principles, the "Data on the Web Best Practices" are being developed (currently a W3C Candidate Recommendation). This set of best practices is intended to improve usability in the Linked Data world by making the Web of Data a stable ecosystem for open data and online publication of research data, with access restrictions, authentication and authorization for sensitive data. They rely on vocabularies and standards and follow the Web architecture principles described above.



One of the key aspects of the framework is that it provides a template for describing each best practice consistently. The template consists of a rationale, an intended outcome, a possible approach to implementation, a way of testing whether the best practice has been met, relevance and benefits. This last criterion classifies the benefits provided by each best practice in eight dimensions:

* Comprehension: provides human-understandable data semantics.
* Processability: supports machines-actionable data manipulation.
* Discoverability: provides metadata to improve search and discovery data content.
* Reuse: maximizes reuse of the data content by different types of consumers.
* Trust: improves confidence in the data provided.
* Linkability: supports link creation between data sets and content.
* Access: support human and machine access in a variety of forms.
* Interoperability: facilitates consensus among data publishers and consumers.

These best practices span a wide variety of topics, from metadata provision (BP1,2,3), provenance (BP5), URI usage (BP9,10,11), standardized formats and representations (BP12,13,14,15), APIs (BP23,24,25,26), and others.

The FAIR principles discussed above are a very good example of the challenges Agencies face in trying to integrate disparate, "fit for use" data sets (and sources) as part of a broader data exploration, search, discovery, and analyze activities.

Entity Resolution for Statistical Production

Identity (or entity) resolution is the identification and linking of different instances/representations of the same real world object. Some sources of entity duplication are:

* Differences in representation across databases and/or frames from multiple sources of admin data (e.g. same business with slightly spelling different in name, address, etc., and entirely different ids)
* Differences in states/versions/variants across multiple repositories (preliminary, final, raw, imputed, v2010, v2012, etc.)

Entity resolution has been originally developed in the data integration domain and has been applied more recently to knowledge representation and the Semantic Web.  
There are two facets to entity resolution: entity management (data/information integration) and ID management (semantics, knowledge representation).

**Entity management**:

1. Resolution: determine when two entities are the same
2. Linkage: link entities that are the same, even when information is not 100% consistent/compatible (e.g. incomplete, erroneous, and out-of-date)
3. Fusion: solve conflicts and merge all records about the same entity into one

Deciding whether records match is often computationally expensive and domain specific. One way of dealing with this problem is to have entirely different approaches for each entity domain, i.e. people, businesses, data sets, metadata, etc. Another one is to divide the problem into a generic approach (for iterating over the data set, deal with constraint and merge/link records) and a set of domain-specific functions (to determine whether two records refer to the same entity). The latter approach allows for a separation of concerns between the data management algorithms and the actual pairwise domain-specific resolution functions.  
Synonyms: de-duplication, record linkage, data fusion.

Some usage within statistical organizations:

* linking statistical records from different admin sources referring to the same entity (person, business, etc)
* linking different versions/states/representations/copies of the same statistical information object (classifications, questionnaire, data sets, etc.) across different repositories/data stores.
* Big data.

**ID management**:

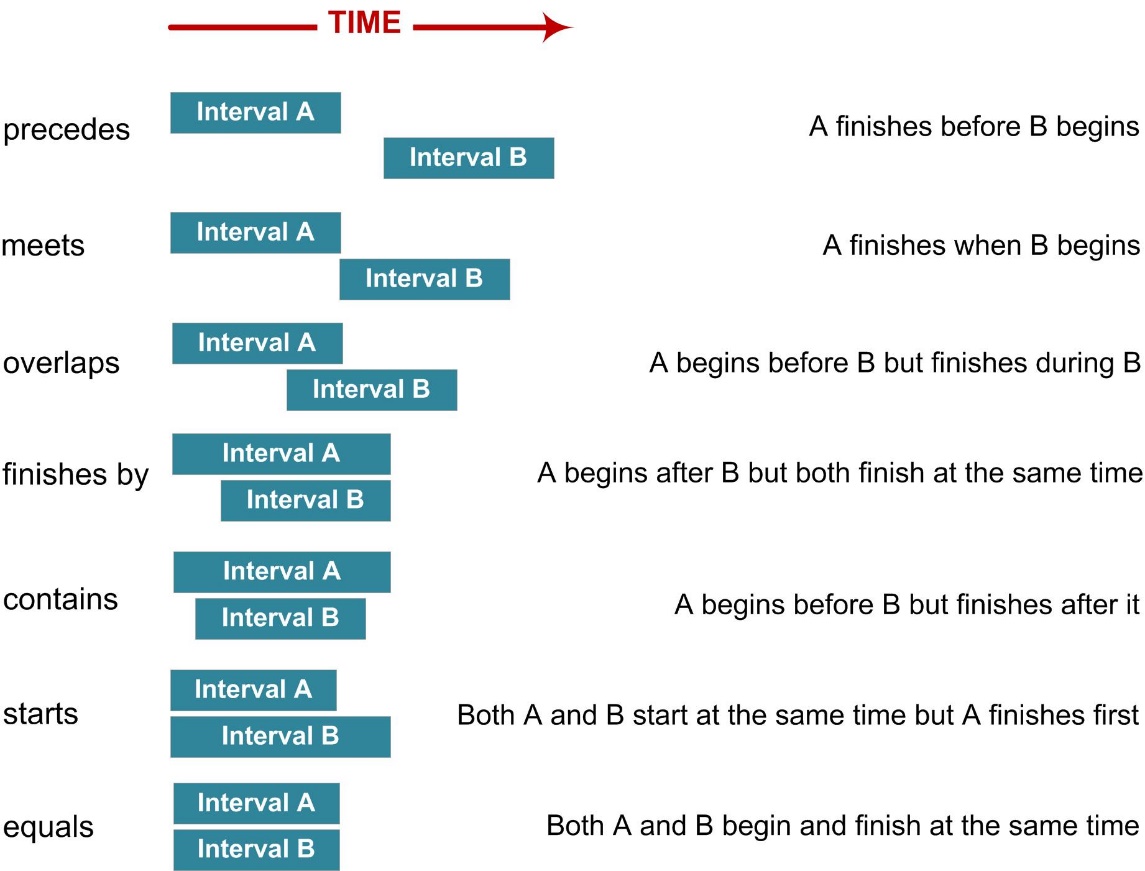
Most statistical organizations face the challenge of managing a set of federated, heterogeneous repositories for their data and metadata assets. Physically centralizing all data and metadata might not be practical in all cases; or even desirable given that centralization at the physical level tends to create systems that are brittle and more resistant to change. In addition, the increasing use of admin data (and big data in the future) in the statistical process is a growing source of heterogeneity and decentralization. The question is then how to identify (and manage) entities in an environment where some enterprise repositories live with a satellite of more or less heterogeneous local data stores to reduce duplication, maximize reuse and facilitate search and discovery of both data and metadata assets.  
One way of addressing the issue is by developing some sort of conceptual id management system, a high-level conceptual integration mechanism (CIM) that will identify (and reference) entities regardless of where they are stored and what physical identifier they were locally assigned. An Entity Name System (ENS) with broader functionality has been recently developed in the context of the Semantic Web [OKKAM]. The idea behind the ENS is to have a global and unique addressing mechanism for identifying and locating entities in the Web, much in the same way a Domain Name System (DNS) works with Uniform Resource Locators (URL) and physical locations on the internet. In the context of the statistical process, A CIM could use a UUID or URN as a global identifiers for a data set and keep track of the different versions, lifecycle stages and even associated metadata (concepts, structures, etc.) each of which could be stored in different physical locations with different ids. This approach also facilitates integration with knowledge management tools and frameworks, such as RDF/OWL. A CIM could be viewed as a capability (implemented perhaps as a dedicated task service) related to other information capabilities like register, update, search, browse, etc.

Temporal Aspects of Data Management

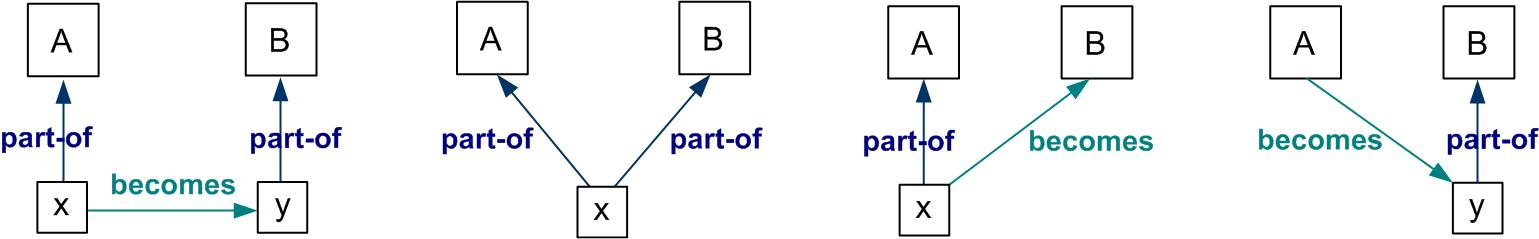
In order to maintain the history of changes an entity goes thru, time intervals can be associated to both entities and their relationships. A time interval is a pair with *start* and *end* instants. The two most common types of intervals are those representing valid time and transaction time. The former is the time in which an entity actually existed in the real world, whereas the latter is the time associated to the persistence of its representation. For instance, a business may have existed for decades (valid time) but the record in a database might have been created yesterday (transaction time).

Valid time intervals associated to entities and relationships provide a basic description of an entity state and its changes, e.g. whether the entity is currently active or not and what are the relationships to other entities that currently hold. They are also used to support time series and longitudinal/historical queries. Most of the literature on temporal data management focuses on managing state changes described by time intervals, e.g. schema evolution and mapping adaptation, temporal extensions for relational and ER models and, more recently, XML and RDF.

Temporal relations can be defined between intervals to express qualitative relationships. In other words, they can be used to define temporal constraints between intervals. The most common ones are the thirteen Allen’s interval relations: twelve asymmetric ones, i.e. *precedes*, *meets*, *overlaps*, *finishes*, *contains*, *starts* and their converses, plus *equals*, which has no converse or, rather, it is the same as its converse. Together these relations are distinct (any pair of definite intervals are described by one and only one of the relations), exhaustive (any pair of definite intervals are described by one of the relations), and qualitative (no numeric time spans are considered). They are defined as follows:

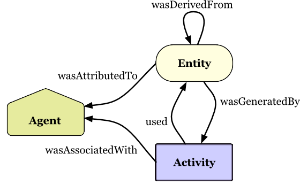


However, time intervals associated to entities and relationships are not enough to explain how entities, and the concepts they represent, evolve over time. Knowing the history of each entity, i.e., how it has been formed and from what, paves the ground for successful entity management solutions and affective information exchange. For instance, a business may be the result of a merger of two businesses, or it can split into multiple businesses after a reorganization. A household may change its constitution with the birth of a child, or it may split into separate households after a divorce. Concepts change as well: the notions of GDP and CPI go thru frequent revisions in terms of what parts of the economy contribute to their calculation. Keeping track of how entities evolved into others (businesses example) and how the parts of an entity change over time (family members and GDP/CPI calculations examples) require types of evolution relationships beyond time intervals. These evolution relationships are based on combinations of two special types of relationships: *part-of*, the usual part-whole relationship from mereology, and *becomes*, which captures the causal relationships, i.e., the interdependency, between two entities. They can be combined in four ways:



These basic evolution relationships can be used as building blocks to model more complex ones, like *splits* (a set of new entities are formed at time *t* with all the parts of a single entity ending at *t*), *joins* (a single new entity is formed at time *t* with all the parts of multiple entities ending at *t*), *detach* (a new entity is formed at time *t* containing a part of an existing entity), and *merge* (a part of an entity ending at time *t* becomes part of an existing entity). Combining these evolution relationships with time intervals and temporal relations (e.g. Allen’s), it’s possible to model complex data and conceptual changes across time beyond traditional time series.

Temporal and evolution relationships can also be extended with additional semantics to model provenance. For instance, in addition to how entities evolve over time we can describe agents and activities associated to the changes to better understand the causal dimension. PROV is a standard for provenance of linked data that provides such a semantic layer. It’s based on the following three building blocks and their relationships:



In this model, the temporal and evolutionary changes described before are caused by agents that induce them via activities. An agent can be a person, an organization, software, or other entities that may be ascribed responsibility. Activities are the dynamic aspects of the world, such as actions, workflows, processes. They generate/change/evolve entities. Agents and Activities can be atomic or composite and the relationships among them, and with Entities, can be specialized to support complex provenance descriptions. Some common specializations include *wasRevisionOf*, *wasPrimarySourceOf* and *wasQuotedFrom* for *wasDerivedFrom*; *startedBy* and *endedBy* for *wasGeneratedBy*. They can be further specialized to explain, for instance, what event produced some merge or split and who or what was the agent behind such an event.

4. Information Capabilities Framework - a reference architecture for analysis

Much work has been done to create information modeling "tools" and standards with a view to facilitating standardized means to describe metadata and data objects and their relationship structures. This allows one to bridge a conversation at a business level (amongst subject-matter experts, practitioners, methodologists, etc.), captured in conceptual terms, to a set of logical and physical representations in the digital world (i.e. technology).

If we look at data management (and by extension, metadata and information management) we can see that there are a set of business capabilities in common that are important for effective acquisition, management, and use of information (data and metadata).

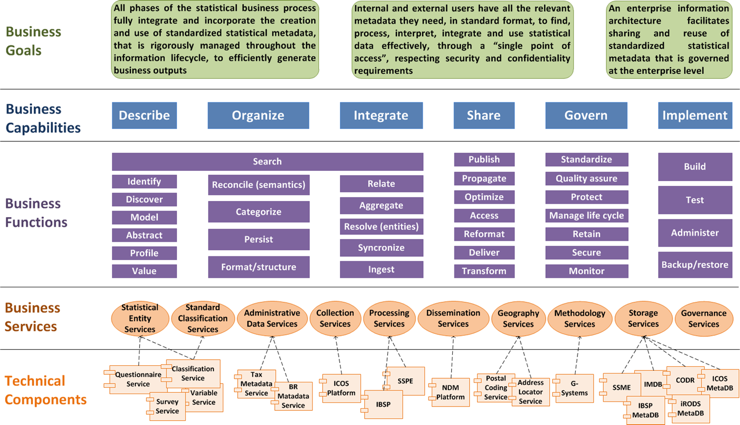
This Enterprise Information Architecture (EIA) view captures key information capabilities in an enterprise architecture, clearly identifying some of the strategic business goals pertaining to information capabilities, and links these to an underlying component and services view. Figure xx shows the framework.

In the “Business Goals” layer one can see an example of statistical data and metadata objectives that align with the "data management” and “metadata-driven” elements in a corporate strategy.

The “Business Capabilities” layer identifies several generic information-related business capabilities – in order to effectively leverage our statistical assets and processes we need to:

* **Describe** – Collect knowledge about information assets: where they are, what format they are in, what level of quality they represent, and their potential value to the enterprise
* **Organize** - Align and structure information assets so that they can be readily found, easily consumed by other capabilities of the platform, and structured in a way that conforms to the organization's standards in regard to syntax (format), semantics (meaning) and terminology (use of common terms)
* **Integrate** - Allow independently designed information structures to be leveraged together toward a common objective
* **Share** - Make data available to consumption points (other systems, end-users)
* **Govern** - Provide for control, levels of consistency, protection, quality assurance, risk assessment and compliance
* **Implement** - Provide the environment for building new capabilities and changing existing ones

It is clear that these are a set of capabilities that are broad and general in their applicability to the creation, management, and use of all kinds of information within an agency – corporate, knowledge, and statistical data and metadata.



In order to provide these business capabilities we need to have a set of underlying functions that support them. The “Business Functions” layer indicates those functions and their alignment with the “Business Capabilities” layer shows how they align. The details of these functions lie beyond the discussion here, however a few comments can help highlight the key functions:

* **Model** – it is important to have a view as to how one identifies and agrees on information objects, how they are described (using GSIM) at a conceptual layer, how in turn they are designed and defined for digital (electronic) representation, and how this then links to underlying solution design, build, and implement functions. Modeling enables “horizontal” agreement between business parties (and technology ones), and it also enables “vertical” alignment from the business (conceptual) layer to the IT (systems – logical, physical) layer, thus improving design effectiveness, quality, and coherence
* **Search** – information assets are not very useful unless they can be “found” and used. An agency should identify its “search” needs and ensure that a holistic approach is used where possible
* **Publish** - while information sets can be produced and consumed within a specific team or survey, agencies are increasingly seeing that “fit for use” datasets are also useful to other survey areas. A Data Management solution may exist whose purpose is to provide a catalogue (register) of these datasets and their associated metadata across an agency
* **Standardize** – in order to ensure that information objects can be produced, managed, and consumed reliably and without ambiguity it is important to develop a means of standardizing both the structure (the objects, their relationships), their representation (the IT-level interchange formats), and their content (semantic harmonization). Standardizing the semantics of information can be very challenging and must be done selectively and carefully. For example, an industry classification scheme such as NAICS represents a standardized structure (the classification tree) and standardized content (the code sets). This is an area where standardized content is needed and desirable. It is generally tempting to set out to standardize all information objects at an agency, however the reality is that this can be an expensive and complex undertaking. In other words, standardization must be driven by the potential value to be received for doing so
* **Manage Lifecycle** – an often-overlooked function is to put in place an effective means of manage the lifecycle of information elements from creation through to mature use, retention, and ultimately disposition and archive. These elements directly link to Information Management (IM) functions that may be managed at a “whole of government” level or individually within an agency. It is important to identify the intended lifecycle for information at its definition stage – how long will we keep this information? Where will we keep it? Will it last forever or will we dispose of it (delete) it? What are the legal or regulatory requirements for lifecycle management?

The “Business Services” layer in the model identifies a number of statistical information services that are commonly found – statistical data and metadata, classifications, administrative data, and registers.

Finally, the “Technical Services” layer identifies the underlying solution components that support and implement the functions and services identified above. In constructing this view we are able to see the links between business goals and strategies, the general capabilities and functions required, the business services, and finally the supporting technical components.

5. Vision - Statistics Netherlands (CBS) Case Study

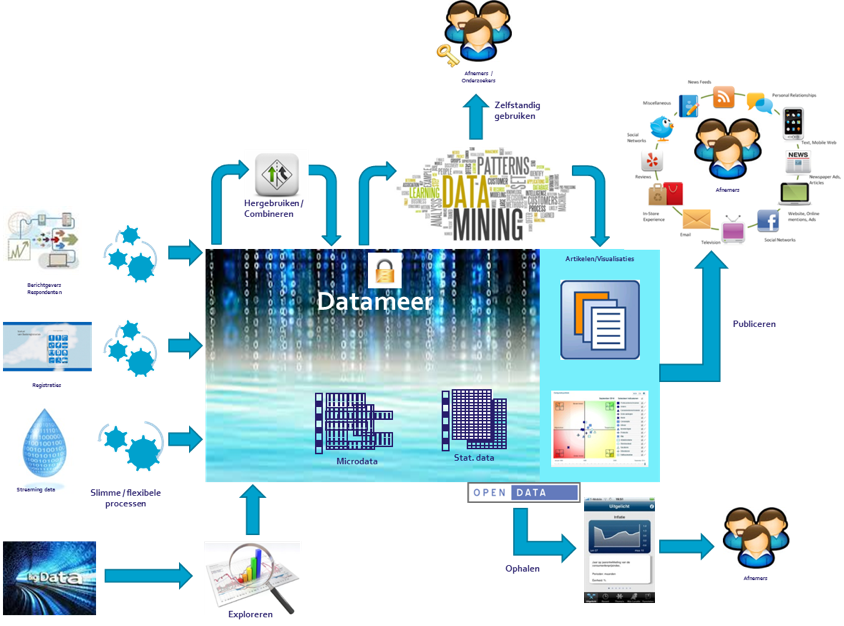
Context

Statistics Netherlands performs direct data collection and has been extensively using secondary data sources including administrative data (for example government registers) and emerging big data (for example road sensors and web-scraped data). This results in a wide variety of input data sources and formats. In the process/analyses phase datasets are stored and managed in some central data stores (functionally similar to Data Warehouse or Data Marts) as well as many decentralized data stores and files within survey processing platforms and systems.

As a result these datasets are described in their own metadata vocabulary and are stored in various formats using different storage techniques, e.g. file base storage using CSV or fixed width, relational database tables, document management systems etc. In order to easily find and use the datasets, Statistics Netherlands decided to investigate and implement a concept, called a ‘Data Lake’.

The main goal of the Data Lake is to empower internal users and analysts at Statistics Netherlands by allowing them to find the datasets that they need for analysis, to access these datasets and to perform operations on them.

A schematic view of the ‘Data Lake’ is represented in the figure 1 below.



Vision

Specific goals from an end-user perspective:

* Enable more phenomenon based output (a phenomenon is a striking event that you want to explain with data)
* Enable more current and coherent statistics
* Stimulate the reuse of data
* Accelerate the statistical processes
* Grow and stimulate the access to a large number of existing and new data sources
* Provide faster response and output to requests from external clients
* Accelerate the design process around collecting and storing data

In its current phase Statistics Netherlands is undertaking research, based on practical user stories to define capabilities and investigate technologies and techniques for next generation data management.

Capabilities:

Below are listed key capabilities that are required to achieve the goals described in the vision. The starting point for the list is that the existing statistical datasets are already stored somewhere, and that these datasets therefore don’t need to be obtained again.

**Data governance:**

* Assure Metadata Quality: each data source must be described with high-quality metadata so users can have all information they need for their use of datasets. For example, statement about the quality and reliability of the data.
* Secure Data Source: this enables authorization of users to get access to data they need.
* Protect Data Source: protect confidentiality of data, such as ability to make certain data unrecognizable (for example Data masking).
* Usage Monitor: this provides monitoring and measures about the use of the data (how often are data sources used, who is accessing them, etc.)

**Metadata management at the enterprise level:**

The content structure and the semantic meaning of statistical data must be recorded. Furthermore, it is possible to trace where data comes from and what path (for example transformations, linkage) has data made before current value (data lineage).

* Manage content types: this concerns the management of names and definitions of object types, such as person, ride, transfer, etc.
* Manage classifications: this concerns the management of classifications and associated code lists.
* Manage variables: this concerns management of names and definitions of variables. These variables are always related to an object type with the relevant unit of measure (number, weight, currency, value, etc.), or link to classification.
* Manage data source descriptions: this concerns management of the (logical) structure of all data sets. This structure describes which object types and variables are in a relevant data set.
* Manage data source relations: datasets can have common object types or variables. This capability enables the management of such relationships.
* Data Sources Register: catalogue with powerful search functionality that enables users to quickly find the desired variables.
* Relate data sources: many data requests cannot be delivered by a single source (dataset), but rather require integration of multiple data sources. This coupling must be facilitated in such a way, that it is clear to the user which fields are eligible for this purpose and the conditions which apply to it.
* Derive Views: enables extraction of new information (often not as a "copy" of a resource, but a selection and / or aggregation or clustering). Merging data provides an added value.
* Standardize: it is important to make arrangements in regards to definition of metadata, quality of data sources, allowable source data formats, etc.

**Logical Architecture layers:**

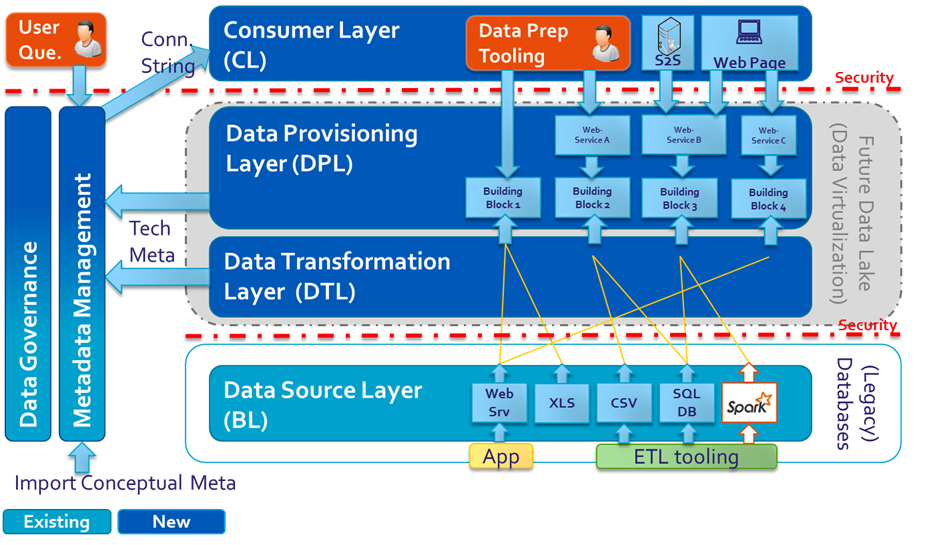
Data Lake architecture has the following logical layers:

1. Data Source layer provides original data to the “Data Lake”. This storage layer could it fact be declared out-of-scope for a data lake project as the virtualization technology does not have any impact on this layer.
2. Transformation Layer enables data transformations, for example data format transformation capability and the capability to create new data sources from already attached data sources. Each data set in the Data Source Layer will in the proposed data virtualization solution have a representative in the Transformation Layer (aka base data sets). There is no data to be found in this layer, only the definition of data sets that will be used.
3. Provisioning Layer makes logical data structure available to users and systems using open standard interfaces and protocols. In comparison with the data sets in the Transformation layer, redundancy in the provided data is allowed (certain objects will be included in multiple data sets – one can find data in de-normalized form). The data sets are designed on the bases of a phenomenon (or topic). Data can be provided as a data set or as a web service, which makes it possible to include the data lake in a Serviced Oriented Architecture (SOA).
4. Consumer Layer consists of tools and systems that actually process or analyse the data. They communicate with the provisioning layer to get access to the chosen logical datasets. Data Preparation, Data Analytics, Data Visualization are sub-domains of this layer.

These layers are accompanied by the cross-cutting layers:

1. Corporate metadata management and
2. Data governance.

A schematic view of all layers is represented in the figure 2



Scope of research

Statistics Netherlands is investigating two key technologies that need to be combined to achieve these goals: a Data Virtualization solution and a semantically enabled Metadata Management system. Data virtualization technology will act as single (logical) data platform (transformation and provisioning layer). Semantic metadata management will provide conceptual model and technical metadata management solution (corporate metadata management layer and part of data governance layer).

Benefits of the new (logical DW) architecture:

* Increased accessibility to available datasets physically stored in different locations and multiple formats without the need for costly data migration and standardisation required to centralise data in single physical data store;
* Logical separation between data sources layer and data consumer layer allows flexibility for example adding (registering) new data sources while protecting analytical and processing systems from change in data sources (ie big data);
* Decoupling of different layers provides mature versioning capability for every data set, meaning a data set can have multiple versions enabling a controlled change management process without time pressure on data set related applications (a timeframe can be provided for the transition from version x to version x+1);
* The data transformation layer can be used to define transformations and other functionality (for example confidentialisation) and make it available across all users/systems reducing development time and fostering reuse;
* Different tools and applications in Data provisioning layer have access to predefined (virtual) datasets, transformations and other common functions (data versioning, data quality, security confidentiality) without the need to copy datasets and replicate these functions in each individual tool or application;
* Faster, phenomenon-based analysis and reporting based on semantic relationships that help find relevant data faster;
* Increased query performance (especially for file based data sources) because of query optimisation and cache-ing methods in transformation & provisioning layer;
* Possibility to add new types of data sources (unstructured, high volume etc.) that require different storage, have various levels of data quality etc.

6. Vision - Statistics Canada Case Study (Metadata Management)

Context

Statistics Canada launched its business transformation journey in 2010 with its Corporate Business Architecture (CBA) transformation initiative. As part of this program, key principles addressing the management of data and the use of metadata-driven approaches. Although we have had a metadata management solution in place for many years, it has become apparent that it is very limited by both its underlying technology as well as its ability to integrate into an enterprise-level data and metadata management architecture. The Picasso project was launched in response to this, with the goals of providing an evolvable platform supporting data and metadata search and discovery based on a powerful core.

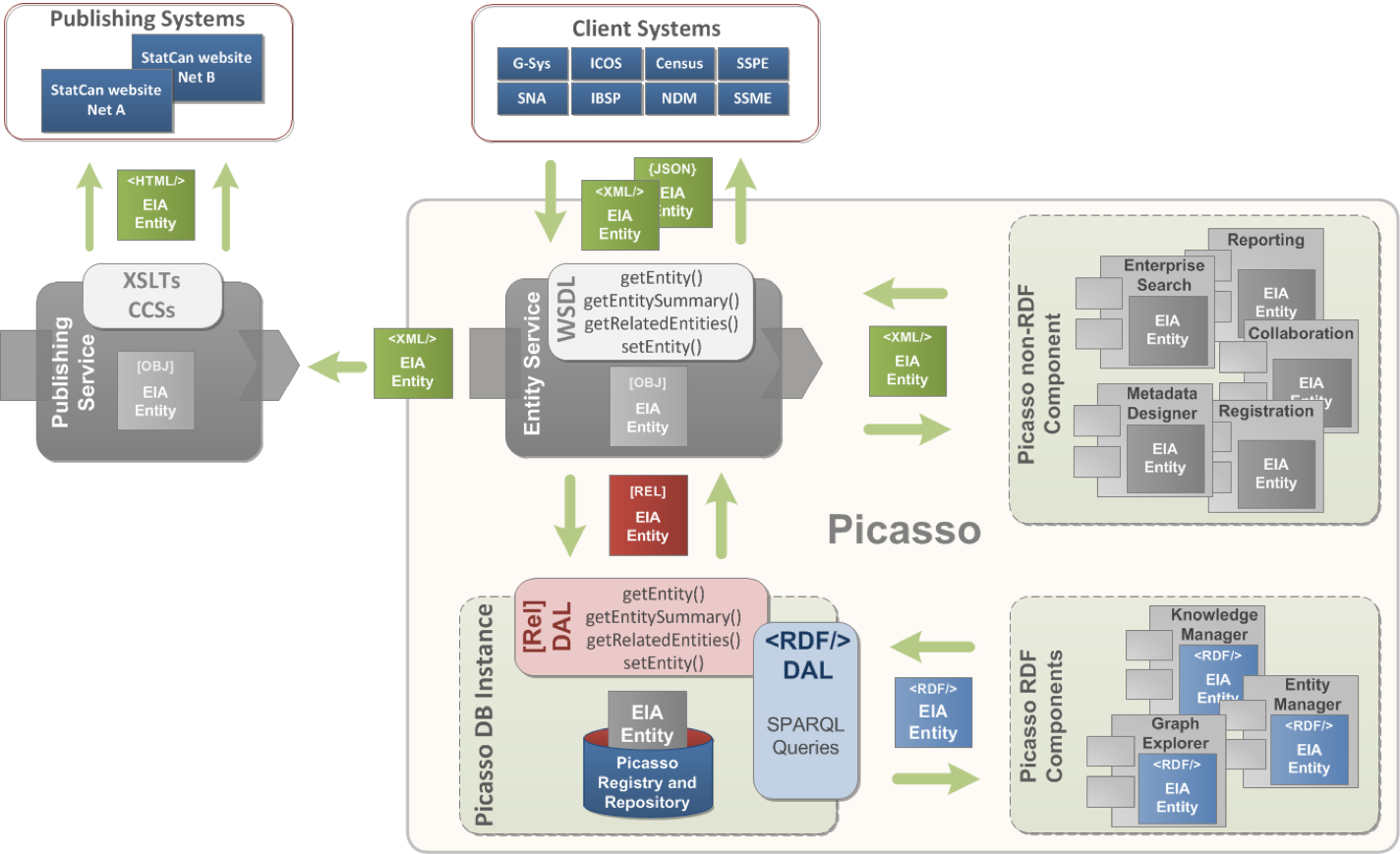
Capabilities

The Picasso vision and solution addresses key capabilities from the Information Capability Framework described above:

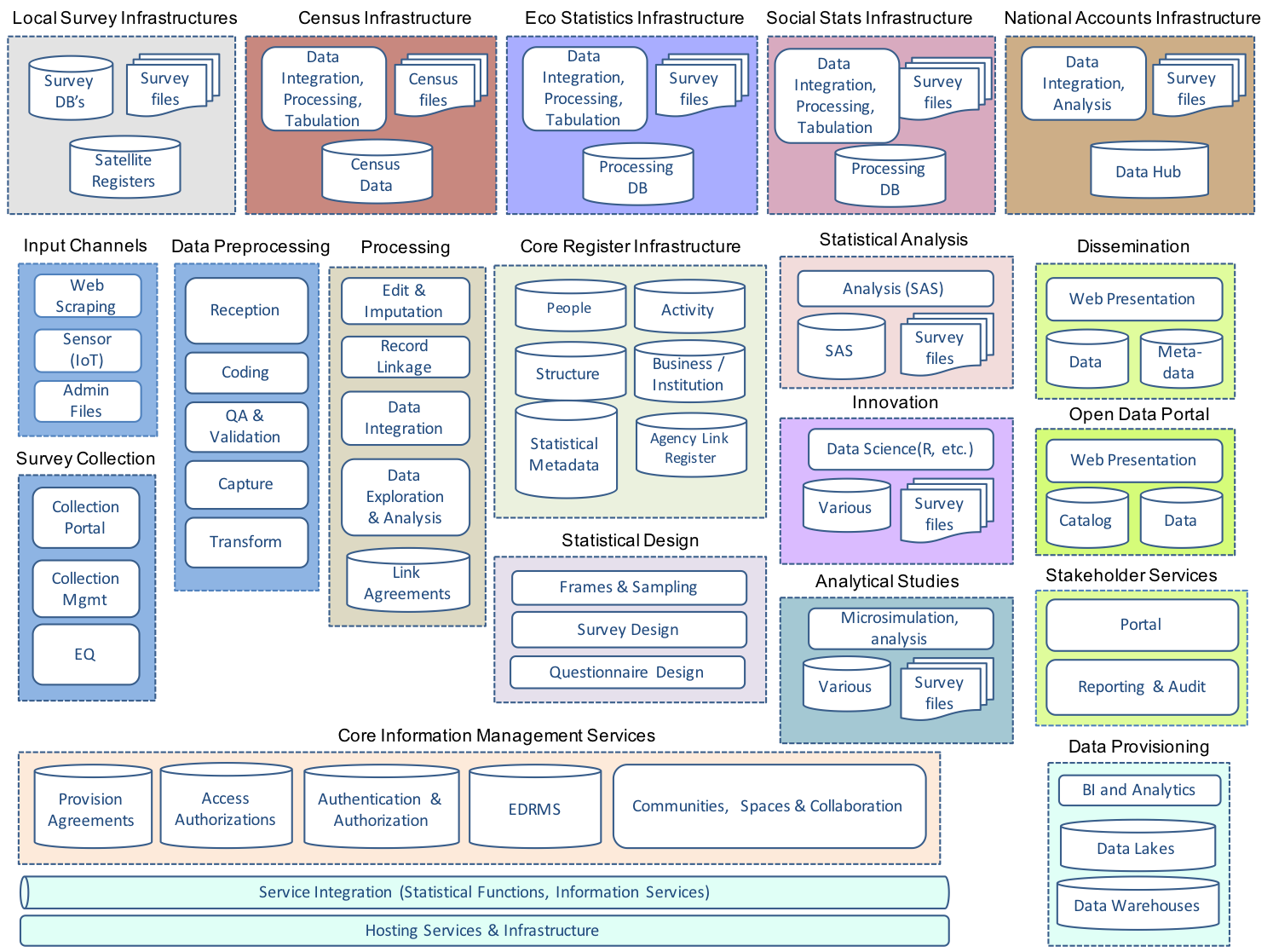
* **Describe** – Providing the focal point for the capture of statistical metadata (including references to fit-for-use data) that is extensible as new directions (e.g. administrative data, Big Data, new sources and products) emerge, with a GSIM-based entity roadmap core, with powerful Search and Discovery functions at an information element (as opposed to container) level
* **Organize** - Leveraging an underlying RDF-based (semantic web) representation and persistance technology
* **Integrate** - Providing services to support the integration of metadata across statistical production, as well as statistical knowledge
* **Share** - Providing standardized services to support the propagation of statistical data and metadata based on common information exchange models (e.g. DDI-based XML exchange)
* **Govern** - Supporting registration and curation workflows in support of Centres of Responsibility focused on what needs to be captured, what structure it has, and (in some cases) standardized semantics (content) in support of business goals
* **Implement** - Delivering a Metadata Core platform based on graph-relational technology, with a data access layer (DAL), metadata services, and extensible browsing, search, and discovery services

Architecture

The following figure illustrates a view of the key components and interfaces of the solution.



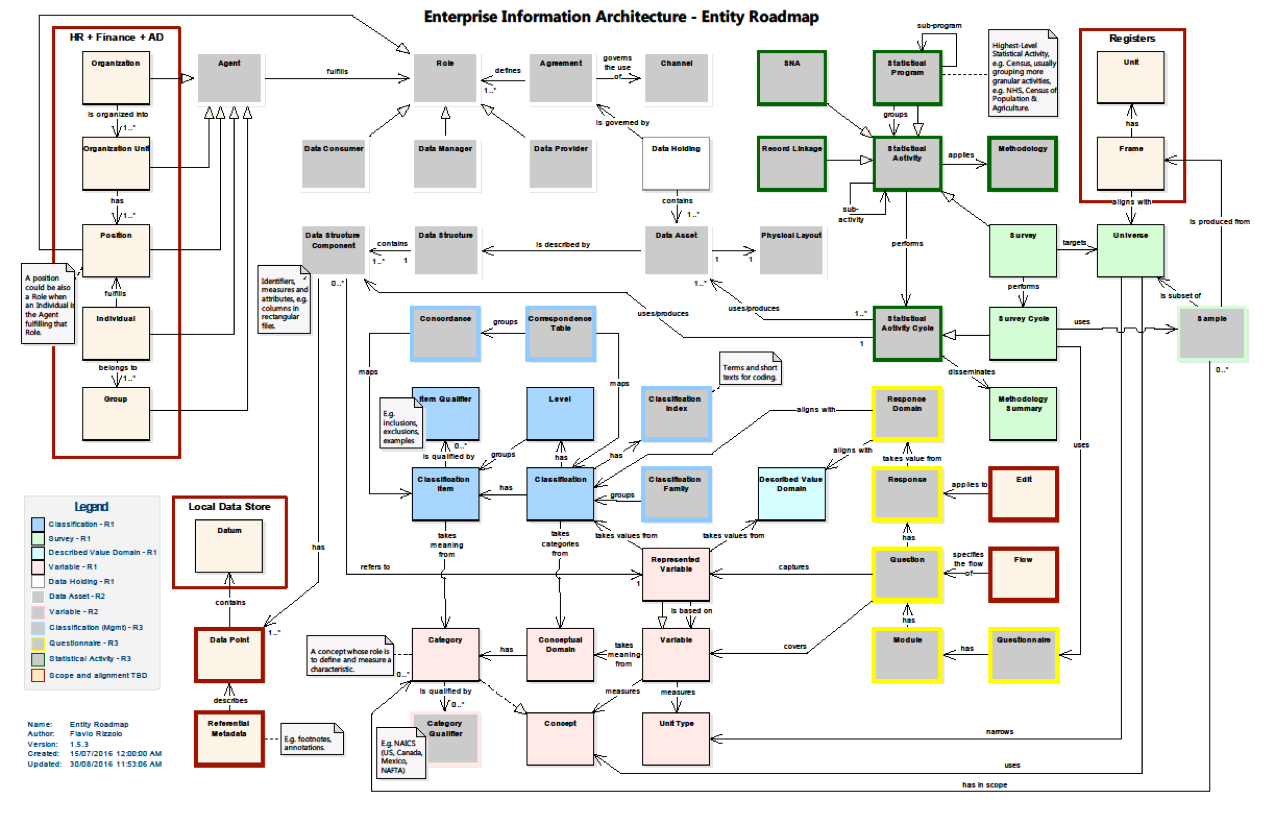
From a core systems contextual view, the following diagram shows a generic view of key functions and services - the reader can see that the statistical metadata management function can be seen as part of a possible core register infrastructure.



One of the challenges is to reconcile the means by which data is managed, shared, reused, and archived in such a picture. In our case, Picasso will provide the statistical metadata management function, and flexibly support a combination of local and centralized data persistence through data asset registration and linkage to metadata elements (such as concept, variable, survey). In addition, we have identified necessary advances in data provisioning technology areas and are engaged in research in those areas.

In looking at the Case Study from the Netherlands (above), one can see that we both have a complementary focus - metadata management and logical data warehouse (data lake, provisioning).

In order to drive agreement on metadata to be captured, its structure, and the delivery roadmap, we have focused on creating and using an Entity Roadmap (below) based on GSIM that provides the supporting tool for discussions with the business and its Centres of Responsibility that ultimately become part of the vision and roadmap for our approach.



At the time of writing, we are actively designing and building our metadata core engine, have initial search and client functionalities in place, and expect a limited rollout in 2017. The core implements RDF and relational approaches to the metadata and data asset registration.

7. Governance and Security

Security and Risk Management

In adopting new techniques to better manage data and metadata in support of greater reuse, agility, experimentation, and new product creation, there is an inherent risk in streamlining the means by which data can be discovered and retrieved. The "security in chaos, it's hard-to-find" attribute disappears, which means that all mechanisms must be robust and subject to audit. With increasing diversity in sources with associated supply agreements, it is essential that agencies meet and demonstrate that they meet their obligations to stakeholders.

There are three key areas to look at:

* Confidentiality - data and confidential metadata must be protected and limited to "need to know" and "authorized to know". This applies to all data provided in confidence, and must also apply to the results of record linkage. Techniques include robust authentication and access control, anonymization of identifying keys, and advances in format preserving encryption (FPE) and other data masking techniques. Care must be taken to ensure accidental linkage results are disclosed that extend beyond what is authorized. The sensitivity of data and metadata must be identified at the start of the discussion, not partway through, to ensure that the necessary controls are in place.
* Integrity - data and metadata must be delivered in a quality fashion to authorized users internally (and possibly through research agreements), and must not be accidentally or intentionally susceptible to modification, either at rest or in transit. This includes malicious agents as well as data transformation and integration technologies and techniques.
* Availability - data and metadata that is more powerfully combined, made available, and shared in support of statistical production, innovation, and experimentation must meet the needs of those activities and processes, not just at a "system availability" level (the traditional view) but also at the level of the data itself.

Techniques to address these elements generally fall into three broad categories of controls - administrative, management, and technical controls. As we apply next generation techniques we must ensure that underlying infrastructure services (such as directory services, access controls, identity and credential management, authentication, and logging, reporting, and auditing) are keeping pace to meet the needs of the future.

Advances in data loss detection and prevention should also be investigated for their ability to aid in this area.

Governance

The rise of new technologies and data types/methods presents a number of related challenges for the management of data. Many of these developments lead to a desire to repurpose (reuse) datasets for uses which were never previously considered. This requires detailed metadata to be stored with the data to allow such repurposing, and for it to be stored in a way which is discoverable and understandable over time. In some cases there may be concerns over the transparency of data use, perhaps requiring renegotiation with data owners, or targets communication with survey respondents/the public.

Other challenges to be overcome in this space is how to ensure that data remains confidential as new technologies become available that make sure that data does not become subject to intentional or spontaneous recognition. Creating statistical linkages keys which will not allow future recognition/reversal can be particularly challenging.

Another complexity is the measurement of quality in new data sources, or using new techniques/technologies. Rarely such processes fit nicely within our current understanding of quality for official statistics, developing new methods or understandings of quality is likely to move much more slowly than technological advances and the desire for information. There are potentially cultural issues that may need to be overcome to ensure the most effective use of these technologies are made.

To ensure that NSOs and related organisations can make the most effective use of the data and technology resources as their disposal, it will be important to ensure that their organisational frameworks allow for this. Such actions would be country specific, but may require changes to legal or access environments, or frameworks and communication tools for data providers.

8. Next Steps

It has not been possible in the time provided to progress more deeply in all areas. The intent of this paper is to share a coherent viewpoint across a number of topics, to describe activities in agencies, and to identify areas for further work and action.

Our study concludes that the need is high for application of new techniques for data management and provisioning. Technology options abound, with rich offerings in both the commercial and open-source marketplaces. HLG-sponsored work has advanced a number of important standards to support integration, description, and use of data. The Data Integration work of 2016 has focused on advancing in the areas of integrating survey and administrative sources, new (and traditional) sources and Big Data, integrating geospatial and statistical information (an important area for visualization), and validating official statistics. This work has moved forward from a methodology perspective, and it is important that the underlying management capabilities keep up.

For the 2017 program year, we see a number of possible activities:

* Leverage the Data Architecture proposal to create a reference architecture for data management
* Possible "blue sky" projects in the area of linked data and linked metadata, building on activities in a variety of agencies
* Review of the current state of GSIM, GSBPM and others for weakness in supporting data management
* Further collaboration with ESSnet activities in relevant areas

9. Acknowledgements

White Paper contributors: Matjaz Jug, Paul Grootens (Netherlands); Keelie Williams (Australia); Keith Chung (New Zealand); Flavio Rizzolo, Robert McLellan (Canada)

Next Generation Data Management working group: Matjaz Jug, Paul Grootens (Netherlands); Keelie Williams (Australia); Keith Chung (New Zealand); Flavio Rizzolo, Robert McLellan (Canada); Janusz Dygaszewicz (Poland).

Reviewers: David Barraclough (OECD); Juan Muñoz (Mexico); Jean-Marc Musseux, Luca Gramaglia (Eurostat).

10. References

Allen, James F.: Maintaining knowledge about temporal intervals. Communications of the ACM (1983).  
Benjelloun O., Garcia-Molina H., Menestrina D., Su Q., Whang S.E., Widom J.: Swoosh: a generic approach to entity resolution. VLDB J (2009)  
Berners-Lee, T. Semantic Web design issues: Linked Data <https://www.w3.org/DesignIssues/LinkedData.html>  
Berners-Lee, T. Uniform Resource Identifier (URI): Generic Syntax, <https://tools.ietf.org/html/rfc3986>  
Bleiholder J., Naumann F.. Data fusion. ACM Computing Surveys (2008).  
Bouquet P., Stoermer H., Bazzanella B.: An Entity Name System (ENS) for the Semantic Web. ESWC (2008).  
DOI handbook <http://www.doi.org/hb.html>  
Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext transfer protocol - http/1.1 (rfc 2616). Request for Comments.<http://www.ietf.org/rfc/rfc2616.txt> (1999).  
FORCE11, The FAIR Principles, <https://www.force11.org/node/6062>  
Gregersen H., Jensen C.S.: Temporal entity-relationship model: a survey. IEEE TKDE (1999)  
Gutiérrez C., Hurtado C.A., Vaisman A.A.: Temporal RDF. ESWC (2005)  
Herzog T.N., Scheuren F.J, Winkler W.E.: Data quality and record linkage techniques. Springer (2007), ISBN 978-0-387-69502-0  
ITU. Recommendation X.667 (10/12). Information technology - Procedures for the operation of object identifier registration authorities: Generation of universally unique identifiers and their use in object identifiers, (2012)  
Lerner B.S.: A model for compound type changes encountered in schema evolution. TODS (2000)  
Moats, R.: Urn syntax. Internet RFC 2141, May 1997  
Rizzolo F., Vaisman A.A.: Temporal XML: modeling, indexing, and query processing. VLDBJ (2008)  
Rizzolo F., Velegrakis Y., Mylopoulos J., Bykau S.: Modeling concept evolution: a historical perspective. ER (2009)  
Soo M.D.: Bibliography on temporal databases. SIGMOD Record (1991)  
Tahmasebi, N., Iofciu, T., Risse, T., Niederee, C., Siberski, W.: Terminology evolution in web archiving: Open issues. International Web Archiving Workshop. (2008)  
Velegrakis Y, Miller R.J., Popa L.: Preserving mapping consistency under schema changes. VLDB J (2004)  
W3C, HTML 5: A vocabulary and associated APIs for HTML and XHTML. W3C Working Draft, August 2009. <http://www.w3.org/TR/2009/WD-html5-20090825/>  
W3C: PROV Model Primer. <https://www.w3.org/TR/prov-primer/>  
W3C, RDF 1.1 Concepts and Abstract Syntax, <https://www.w3.org/TR/rdf11-concepts/>  
W3C, RDF Schema 1.1, <https://www.w3.org/TR/rdf-schema/>  
W3C, Data on the Web Best Practices, <https://www.w3.org/TR/dwbp/>  
W3C, Architecture of the World Wide Web, Volume One, <https://www.w3.org/TR/webarch/>

INSEE presentation at Standards Workshop, September 2016