TNO report

|  |  |  |
| --- | --- | --- |
| TNO 2019 R12001 | | |
| Selection of a data storage, management and analysis ecosystem for SR | |
|  | | |
| Date | 17 December 2019 | |
|  |  | |
| Author(s) | Dr. Ir. H.M. Böhms  Dr. R. Sebastian  Dr. Ir. G.T. Luiten | |

|  |  |
| --- | --- |
| Number of pages | 32 (incl. appendices) |
| Number of appendices | 1 |
| Sponsor | VP Buildings and Infrastructure |
| Project name | VP-INFRA 2019 - Duurzame Infrastructuur en Mobiliteit, WP5 Locating and sharing data |
| Project number | 060.37826/01.06 |

All rights reserved.

No part of this publication may be reproduced and/or published by print, photoprint, microfilm or any other means without the previous written consent of TNO.

In case this report was drafted on instructions, the rights and obligations of contracting parties are subject to either the General Terms and Conditions for commissions to TNO, or the relevant agreement concluded between the contracting parties. Submitting the report for inspection to parties who have a direct interest is permitted.

© 2019 TNO

Summary

Data plays an important role in the research and consultancy work as done by the TNO Structural Reliability (SR) expertise group. There is a serious need to standardize the ways we manage data so that we can (re)use it more effectively, efficiently and easily, both in and over projects.

Therefore, there is an urgent need for a common data management and analysis ecosystem. In this report we propose a feasible solution for (a part of) such a common ecosystem in four steps:

1. Identifying the data management and data analysis requirements in a research environment, such as TNO SR.
2. Sketching a conceptual architecture of the SR data ecosystem.
3. Giving an overview of the state of the art for tools supporting data management and data analysis in TNO SR and pre-selecting the most viable options for SR.
4. Drawing the conclusion on the feasibility of a data management ecosystem, and proposing a possible solution for the research purpose and environment of TNO expertise group SR.

The main conclusion from this study is that there is no single existing data ecosystem or platform solution that fully covers all the needs of SR within the given boundary conditions. Therefore, the most feasible way forward is a combined use of (parts of) several solutions or platforms. The positioning and coverage of these part-solutions as part of the combined ecosystem have been mapped on the SR conceptual IT architecture, which is explained in this report.

As follow-up of this study, three main recommendations are given:

* The selected combined solutions should be subject to an in-depth assessment by applying it in SR’s research projects.
* The integration approach over the selected combined solutions should be further developed.
* The combined solutions should be scaled up beyond SR, especially related to its potential role within the Digital Twin concept.

Contents

Summary 2

Contents 3

1 Introduction 4

1.1 Rationale and objective of this project 4

1.2 Context and scope of this project 5

1.3 Structure of the research project and this report 7

2 Requirements for the data management ecosystem 8

2.1 Information sources from which the requirements are derived 8

2.2 Context and purpose of the intended data ecosystem 8

2.3 Selected focus for defining the requirements 8

2.4 General requirements and fundamental principles 9

2.5 Extended requirements for follow-up assessment of the preliminary selection 11

3 Conceptual architecture of the SR data ecosystem 12

4 State-of-the-Art and preselection 14

4.1 State-of-the-Art of data & tool ecosystems 15

4.2 Preselection of solutions for SR data ecosystem 16

5 Conclusions and recommendations 26

5.1 Conclusions 26

5.2 Recommendations 27

6 Signature 28

Appendix 1: Benchmarking Existing Smart Cities Platforms 29

# Introduction

## Rationale and objective of this project

Rationale:

Data empowers people to make better decisions, both operationally and tactically.

Data is input for data analysis by software functionalities that derive even more meaningful data. Hence we see the following logic:

**Improved data management & data analysis > better data > better decisions**

Everyone in the expertise group Structural Reliability (SR) works with asset-related data. Usually they utilize project specific solutions for data management and processing, which depend on individual projects or individual researchers. Such a project specific approach has led to;

* the practice of reinventing the wheel for every project/researcher,
* an inefficient effort, as more time and effort are often needed for data acquisition and data management compared to conducting technical analysis,
* the lack of re-use of data and knowledge, as there is no overarching data over various projects which prevents the accumulated knowledge to be disseminated and used in other projects, other than via the involved researchers.

Since data plays an important role as the basis for research and consultancy work by the SR department, there is a serious need to standardize the ways we manage data so that we can (re)use it more effectively, efficiently and easily; for example, having the ability to combine data created/captured in different projects over time, which contribute to the body of knowledge in the SR department. Therefore, there is an urgent need for a common data management and analysis ecosystem.

The term *‘ecosystem’* indicates a flexible collection of collaborating ICT tools to create/capture, store, retrieve, manage and analyse data. These comprise databases, computer applications/services, file management systems, data structures, etc. This term also indicates that software tools are not necessarily obtained from a single IT provider or operated on a single platform. In this sense, the flexibility to select and use tools from multiple vendors is considered essential.

The objective of this research is:

The study in this report is performed to tackle a research problem that can be summarised in the Figure 1. The goal of this study is to propose a solution for a data ecosystem that facilitates and supports research activities by TNO’s expertise group Structural Reliability (SR). The intended solution should serve as a ‘bridge’ between the data sources and functionalities. To this end, the research aims to propose a feasible solution for (a part of) a common ecosystem for data management and analysis within SR.

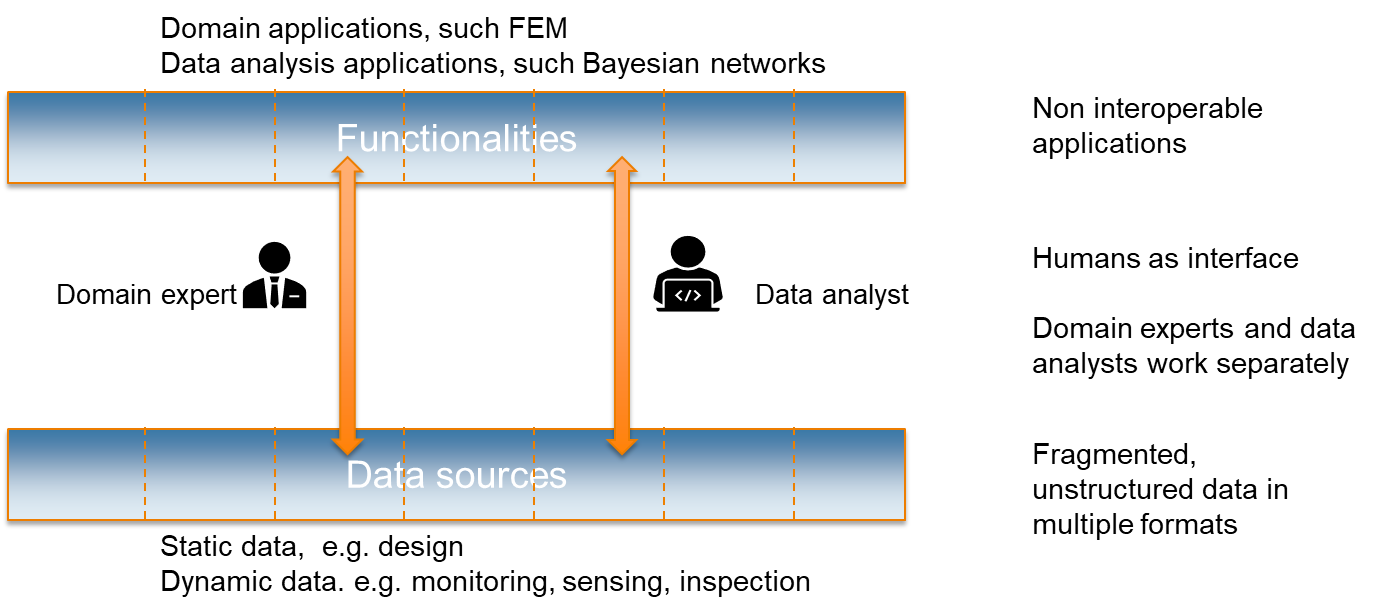


Figure 1: Summary of the research problem dealt with in this study

## Context and scope of this project

Context

In this project, the TNO expertise group of Structural Reliability (SR) is taken as the context for which the feasibility of a data ecosystem will be assessed. In the feasibility study both the added value compared to the costs and the realizability in a research environment are assessed.

For the time being, the focus is put on infrastructure-related asset data, like roads and structures such as bridges, their parts and materials. Buildings, energy, maritime/offshore structures are not the focus of this investigation. Most likely, many lessons learned from infrastructure cases can also be applicable for other TNO research projects, and vice versa.

The ecosystem for data management and analysis corresponds strongly with the concept of Digital Twins, in which data about an asset is gathered and combined with the software functionalities to create/capture, manage and analyse this data. Therefore, the ecosystem that combines data management and analysis is called a *‘Digital Twin ecosystem’* within the context of this study.

Scope

The scope of this project covers the entire Data Lifecycle, i.e.:

* Capture (AS-IS situations)

‘Capturing’ or data acquisition covers collection of monitoring data on the actual behaviours of the asset through sensors or manual inspections as well as 3D (spatial) scanning of the asset’s geometries, damages, etc.

* Create (TO-BE situations)

‘Creation’ is related to defining requirements and Key Performance Indicators (KPIs) for new asset design/construction or for maintenance/improvement of the existing assets. ‘Creation’ provides input for the design and engineering processes through which solutions are developed towards a TO-BE situation.

* Calculate & Analyse (AS-IS & TO-BE)

In ‘calculate & analyse’, the captured existing situation and the proposed new situation are assessed via computational and simulation models ranging from detailed technical analyses to general evaluation of the KPIs.

* Compare & Decide

The captured/created and analysed data is compared, ranked, visualized, etc. in order to support an optimal decision-making.

* Control (operational) & Learn (tactical)

Based on decisions, measures are taken in the real world (either manually or automatically via a control system/actuator). Such measures can include the improvements based on the outcomes from the applied calculation/simulation methods. Usually additional TO-BE solutions variants are created, and more data on the existing situation as part of ‘learn’ alongside ‘control’.

This scope is visualized in a Digital Twin concept in the next scheme. The data ecosystem is an important component in this concept.

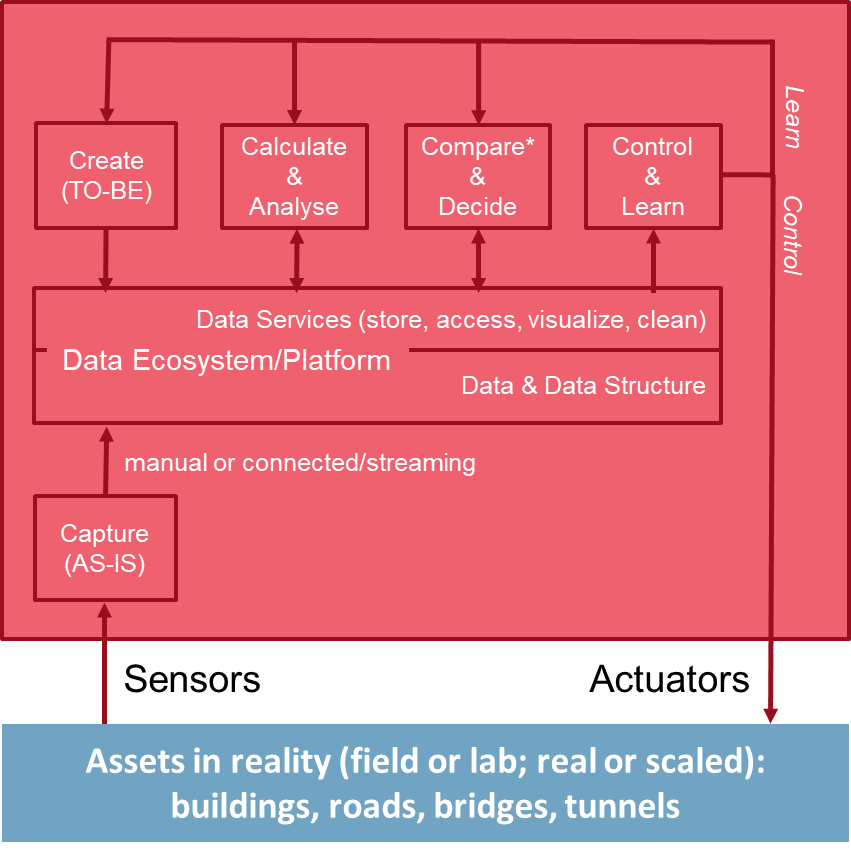


Figure 2: Positioning Data Ecosystem in a Digital Twin concept

The data ecosystem is the place where all functionalities (traditional software applications as well as Artificial Intelligence / Machine Learning (AI/ML) applications) come together to address various data-related aspects, i.e.:

* ‘What’ – data semantics (related to the data structures):
  + AS-IS and TO-BE data.
  + Static and Dynamic (real-time) data.
  + Structured, computer-interpretable data and data for human interpretation (for instance: readable documents).
  + Data structured according to generic national or international standards and data structure according to specific agreements.
  + Meta-data.
  + All data linked together.
* ‘How’ – data formats and data access:
  + Upload/download, data exchange.
  + Direct access.
  + Data sharing.
* ‘Where’ – location of the data and services:
  + Central / decentral.
  + Manual upload/download and streamed from measurement systems (live data).

The data as a part of the ecosystem comprises :

* + Documents.
  + Data sets and data structures.
  + Links between the documents, data sets and data structures.

## Structure of the research project and this report

This project is structured in four main tasks as follows:

* Task 1 identifies the data management and data analysis requirements from TNO SR (see chapter 2).
* Task 2 sketches a conceptual architecture for SR’s data ecosystem (see

chapter 3).

* Task 3 gives an overview of the state of the art for tools supporting data management and data analysis and gives a pre-selection of viable options for TNO SR (see chapter 4).
* Task 4 draws the conclusion on the feasibility of a data management ecosystem and proposes a possible solution for the research purpose and environment of TNO SR department (chapter 5).

# Requirements for the data management ecosystem

## Information sources from which the requirements are derived

The following information sources have been used to derive the requirements for the data ecosystem for SR:

* VP-Infra 2017 – Kansen voor data.
* KIP S&S 2018 – WP5 Data Management.
* KIP S&S 2018 – WP6 Data analyse.
* ERP DT-SI 2019, parts A, B, C.
* KIP BTIC Digitalisation 2019 – WP3 ALIM and Digital Twin.
* Interviews with users of data management and data analysis tools.
* Scientific research data management in scientific domains.

## Context and purpose of the intended data ecosystem

Research in SR has the following general characteristics that need to be supported by the intended data ecosystem:

* On the level of individual project, the data ecosystem should support: project-specific data collected from multiple sources; data processing and analysis focusing on project-specific objectives; the use of tools specifically chosen for the (type of) project.
* On the level of a series of projects, the data ecosystem should support: reuse of data, data analysis results and tools from the previous projects; generalization of data to gain new insights which may result in new knowledge.
* For the purpose of new knowledge generation, the data ecosystem should support: a combination of the data available in SR, new data from other sources, and the domain knowledge in order to create new insights.
* For the purpose of working together with external partners: facilitating collaboration within EU project consortiums, national programmes (such as AsfaltImpuls) as well as collaboration with recurring clients (such as RWS).
* For the wider and longer-term perspective: Digitalization is an actual topic in built environment research. There is an ongoing discussion on the ‘Digitaal Stelsel Gebouwde Omgeving’ (DSGO) that addresses how to structure asset data according to sectoral standards. TNO plays an important role in this endeavour and advises the stakeholders in building and infrastructure sectors. Therefore, it would make sense if we can apply a similar approach for this issue based on the intended SR data management concept.

## Selected focus for defining the requirements

Within the available time and resources, and taking the highly prioritised need into account, the selected focus of this research project is put on the need for a data ecosystem for civil infrastructure research activities in SR.

The majority of SR research focuses on structural reliability of existing assets/structures in the civil infrastructure sector. In such research, SR applies: the knowledge of structural engineering (within the disciplines of road engineering, concrete and steel structural engineering); the data from the design and structural health monitoring of structures, originating from the asset owners, inspections and monitoring; and the tools to analyse structural integrity in order to advise asset owners on the risk, maintenance and lifecycle of their civil infrastructure assets. The hybrid characteristic of research in SR, in which analytical/computational models based on the technical domain knowledge are combined with advanced data analysis, has become the unique competence of SR.

Civil infrastructure assets have several unique characteristics, among others:

* there is often a ‘network level’ on which the individual infrastructure assets are connected together and have their impact on usage;
* real-time actions is not the first priority (unless in case of emergency) as most measures are concerned with long-term planning;
* civil infrastructure assets have a different Level of Details (LoD) below that of the infrastructure network; therefore, the intended data ecosystem should support both LoDs, including the relations between these LoD;
* at present, SR mostly works on existing infrastructures although we address the full asset lifecycle covering the design, build & operation phases.

## General requirements and fundamental principles

This research makes a preliminary selection of the SR data ecosystem based on consolidated general requirements and the fundamental principles as follows:

* SR research effort should not be hampered by the limitations of data and application ecosystems. For instance, in short term, the selected data ecosystem should be able to accommodate the monitoring data from Bridge 705, online streaming from the ongoing Digital Twin demonstrators, real time analysis, programming in python, etc.

The ecosystem should be able to handle raw data, processed data, and analysis. More specifically, the ecosystem should be able to manage the data retrieved from first-hand sources (e.g. streaming from sensors); the data contained in heterogenous files; the data which is retrievable, traceable and identifiable with meta-data; the data which is enriched with semantics; and the data which is processed for analysis.

* There should be a separation between data and software functionality, i.e. data should be independent from the software applications so it can be reused by others through other applications over time.

The data ecosystem should support heterogenous data files: linked data, other standards, CSV files, unstructured documents. It should be open for multiple database technologies and not be limited by vendor lock-in.

The researchers should keep their freedom to use their preferred applications while harmonization and collective learning are stimulated.

* The data ecosystem should be aligned with TNO’s internal IT policies with regards to security and data privacy, the use of SharePoint, GitLab, etc.

It should be easy to use for non-IT experts and cost-effective for TNO.

* The data ecosystem should also be usable for collaborative projects with recurring customers and within multi-stakeholder projects such as EU-projects.

In addition, it should foster collaboration, joint projects and cross-learning between SR researchers, TNO colleagues and external partners.

As fundamental principles, the TNO’s three principles for data management in the infrastructure sector are kept and promoted. These principles are:

1. Use open standards: linked data / semantic web as described by World Wide Web Consortium (W3C) for the technological parts; and existing international standards for the semantics of data structures (sometimes called object type libraries).
2. Follow modelling and linking guide (CEDR-INTERLINK) or NTA 8035 (to appear January 2020) for networks of ontologies.
3. Follow Transportation Network model of INSPIRE for modelling of physical networks, such as roads, which serve as the connecting hub in the network of OTLs. Find similar generally accepted standards as hubs for other asset types, such as civil structures or buildings.

In particular for structured data, the FAIR principles (Findable, Accessible, Interoperable, Reusable) serve as a guideline. The FAIR principles are[[1]](#footnote-2):

* To be Findable:
  1. F1. (meta)data are assigned a globally unique and eternally persistent identifier.
  2. F2. data are described with rich metadata.
  3. F3. (meta)data are registered or indexed in a searchable resource.
  4. F4. metadata specify the data identifier.
* To be Accessible:
  1. A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
  2. A1.1 the protocol is open, free, and universally implementable.
  3. A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
  4. A2 metadata are accessible, even when the data are no longer available.
* To be Interoperable:
  1. I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  2. I2. (meta)data use vocabularies that follow FAIR principles.
  3. I3. (meta)data include qualified references to other (meta)data.
* To be Re-usable:
  1. R1. meta(data) have a plurality of accurate and relevant attributes.
  2. R1.1. (meta)data are released with a clear and accessible data usage license.
  3. R1.2. (meta)data are associated with their provenance.
  4. R1.3. (meta)data meet domain-relevant community standards.

## Extended requirements for follow-up assessment of the preliminary selection

In the report, a full range of requirements are listed, yet the current research is focused on the general requirements presented in the previous section. The extended requirements are described in this section. In follow-up research, a further assessment will be conducted on the preliminary selected data ecosystems. The fulfilment of the functional and technical requirements for the intended data ecosystems will become part of this further assessment. These requirements are concerned with the following aspects:

* Facilitating data services and data analysis:
  + Data Services (document / data management), including:
    - * authentication/authorization,
      * documents & data (content, meta-data & data structure) storage & access (browsing or machine-based),
      * document/data modelling & linking (adding “semantics”),
      * meta-data assignment (including version management),
      * data cleaning like corrections & alignment,
      * data transformation (translation & conversion),
      * data visualization,
      * backup/archiving,
  + Data analysis (data processing & expert computation).
* Adequate level of performance in terms of capacity, reliability and long-term support.
* Reasonable costs, including initial cost, licenses, usage-dependent fees.
* Choices of commercial-on-the-shelf (COTS) and open-source software (OSS).
* Interoperability to support W3C Linked Data / Semantic Web standards and Big Data open standards.
* Compliance to the relevant systems and software used within TNO, e.g. SharePoint, MATLAB, GitLab, Python.
* Flexibility to future adaptations and developments, such as the development of creation/capture, calculation/analysis, decision support, learning & control methods; the combination of different Levels of Detail (LoDs); and the implementation of Artificial Intelligence (AI), especially Machine Learning (ML).
* Alignment with future strategies regarding a Common Data Environment (CDE). A CDE is simply a collaborative environment that everyone uses, following the guidance given under PAS1192 and BS1192, to coordinate information with supply chain members on the project. In summary: The Common Data Environment (CDE) is the single source of information for the project, used to collect, manage and disseminate documentation, the graphical model and non-graphical data for the whole project team (i.e. all project information whether created in a BIM environment or in a conventional data format). Creating this single source of information facilitates collaboration between project team members and helps avoid duplication and mistakes.

(ref:https://blogs.oracle.com/construction-engineering/common-data-environment-cde-tutorial).

# Conceptual architecture of the SR data ecosystem

Based on the aforementioned general requirements and fundamental principles, the conceptual architecture of the intended SR data ecosystem can be shown schematically as follows.

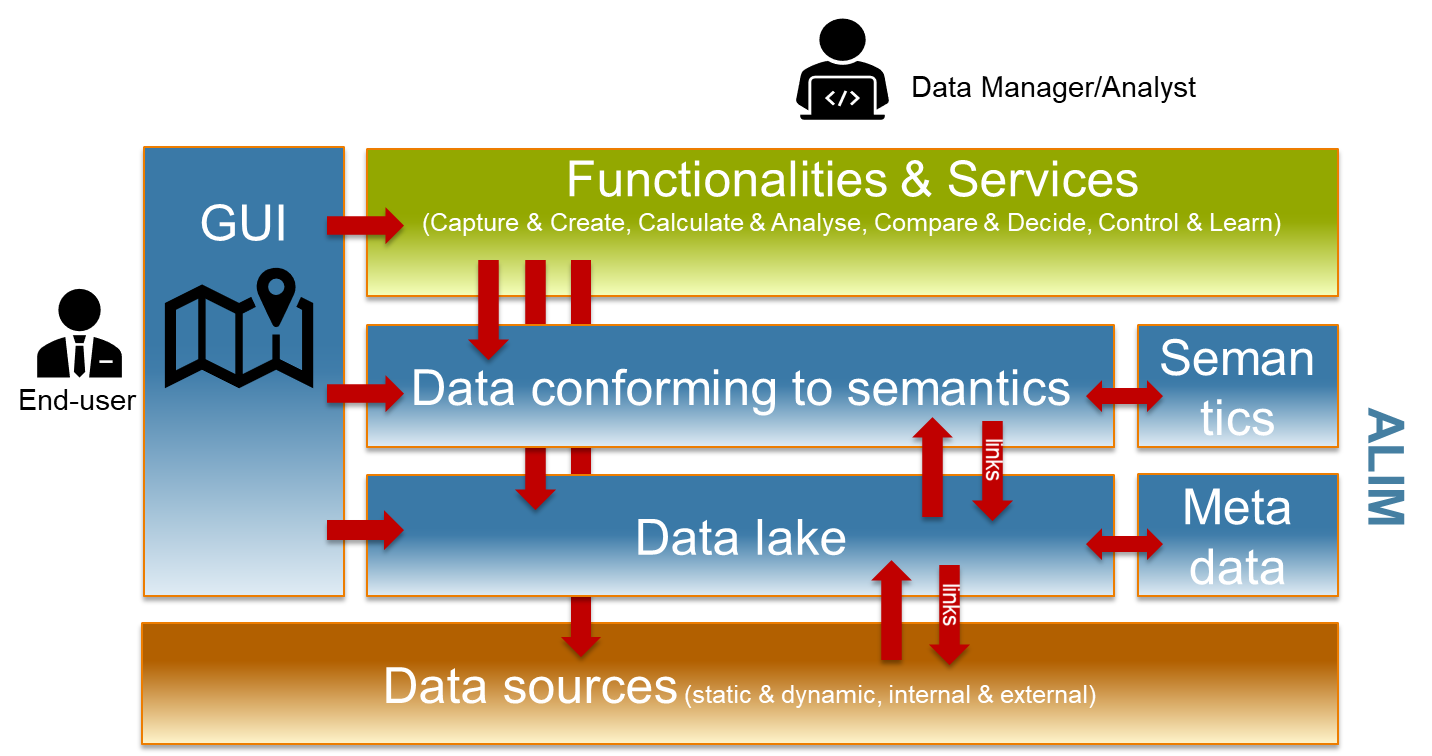


Figure 3: Conceptual architecture of SR’s data management and analysis ecosystem

In general, the orange block represents raw data that can be used for different purposes; the blue blocks represent the IT infrastructure for processing and managing the data; and the green block represents specific tools or methods which are used by the domain experts in a single or multiple projects. Altogether they represent the IT ecosystem for SR. The overall context is Asset Lifecycle Information Management (ALIM) that facilitates analyses in various technical domains regarding building, infrastructure, maritime and offshore assets.

Data sources can be anything. Any document, structured or unstructured provided by any party in any format. It can cover relatively static object data or dynamic monitoring data. Examples are: pdf files, step files, xml files, ontologies in owl, excel workbooks, word contracts, comma separated value files or Revit native design files. For a specific situation (for instance for the client as a specific stakeholder), a selection can be made in the data lake. The transfer between data sources and the data lake selection can take several forms:

* via import/export (over the web: download/upload),
* via links or direct access without copies utilizing URIs, APIs, query languages etc., or,
* via direct real-time streaming of data from a measurement system involving sensors and/or actuators in reality.

Once this data is in the “data lake” it can be enriched/annotated by adding meta-data reflecting its source, status, precision etc. Typically it is also given some explicit context like a position in time and space often already implicitly available in its original folder/file name. For some of these data sets the meaning or semantics can be made explicit in schemas, ontologies etc. depending on the actual formats. All this data, data sources, annotated data, semantic data, can be accessed by services and functionalities depending on the technologies involved.

On the technology side, W3C Linked Data technology –where data is in standard triple (RDF) forms and semantics is provided by ontologies– is recommended. As such, non-linked data can be referenced and all data can be directly accessed via SPARQL or frontends like GraphQL-LD.

Refer to this IT architecture, the state-of-the-art and preselection are described in the next chapter.

# State-of-the-Art and preselection

The approach to the state-of-the-art explorative has been inspired by the framework presented by Peter Nooren (TNO) that promotes a complete and consistent analysis of key platform aspects as illustrated in Figure 4.

For the sake of practicality and feasibility in this current study, this framework is adjusted to the purpose to seek a data ecosystem for SR. The main considerations that are mapped on the original framework are presented in Figure 5.

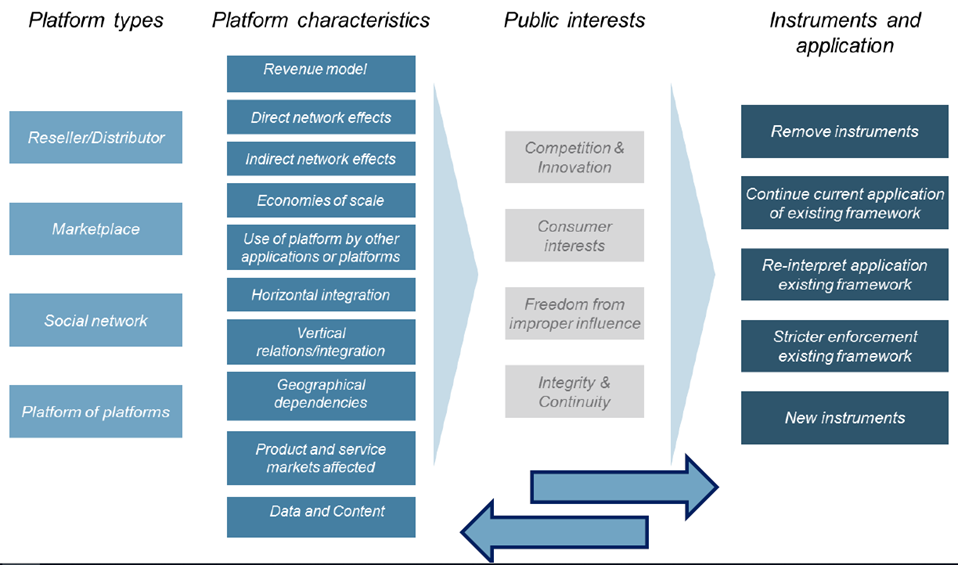


Figure 4: General framework for analysing and selecting IT platforms (Peter Nooren, TNO)

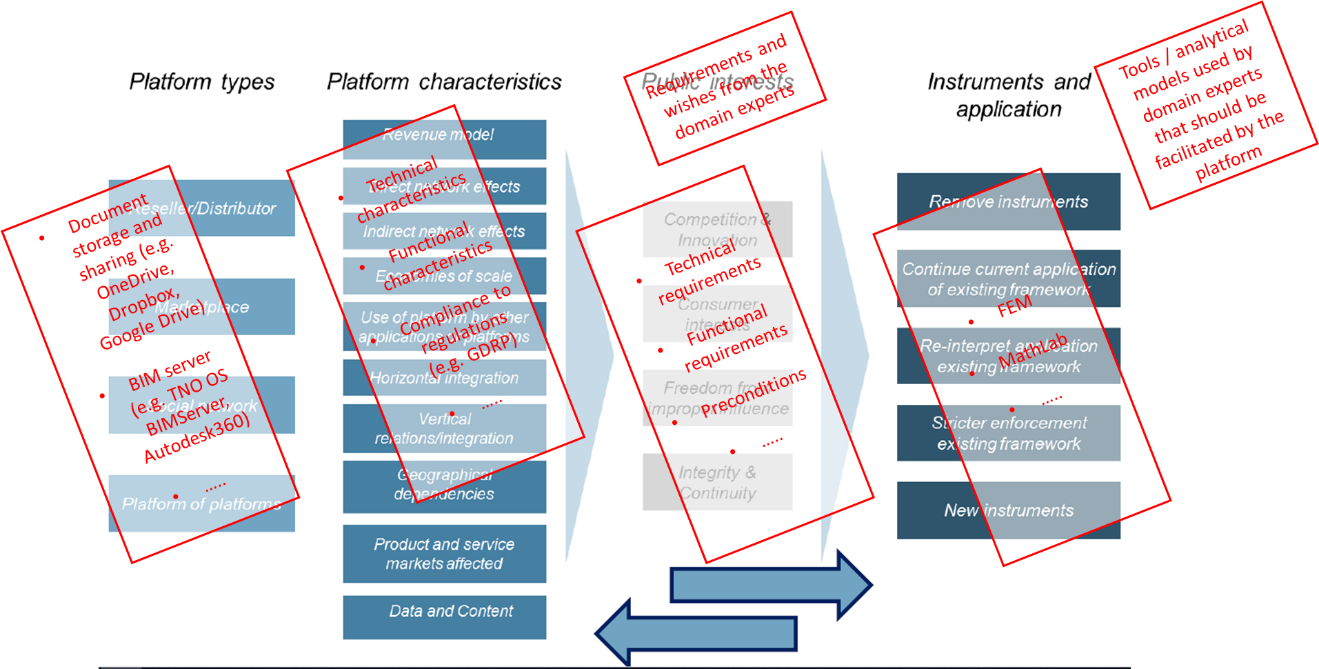


Figure 5: Simplified framework for preselecting SR data ecosystem

## State-of-the-Art of data & tool ecosystems

At present there is a growing number of data and tool platforms / ecosystems, both for commercial as well as for research use. For the explorative study refer to the purposes for SR and the feasibility reasons, the available solutions are categorized depending on their primary focus for:

* Document management.
* Storage of measurement (real-time) data.
* Database management system (DBMS) incl. BIM data.
* BIM and GIS servers.
* Enterprise Data Governance (EDG) based on Linked Data.
* Integrated Development Environments (IDE) for data and functionalities, incl. AI developments.
* Existing solutions/tools developed by/within TNO BI&M.

Within each category, a short list of commonly known/used solutions are presented:

* Document management;
  + - Microsoft SharePoint (standard solution used at TNO).
    - File servers (various solutions available).
* Storage of measurement (real-time) data ;
  + - OSIsoft PI (currently used for our research on Digital Twin of steel bridges).
    - SkySpark (currently used for our research on Digital Twin of buildings).
    - InfluxDB (back-end) / Graphana (front-end) (currently used for research at Department of SD by a.o. Thomas van Dijk).
    - FIWARE/IDS (Industrial Data Space).
    - Siemens, Cisco, Google, Amazon, Microsoft IoT platforms.
    - Various platforms for Smart Cities (see Appendix 1 for the comparison overview refer to research by Yacine Rezgui, Sylvain Kubicki and

Alain Zarli).

* Database Management System (DBMS);
  + - Open Source PostgreSQL.
    - Stardog (currently used in the EU project BIM4Ren).
    - Oracle (incl. the variants with relational and semantic support).
    - Neo4J (property graph based).
    - GraphDB.
    - Casandra (considered for future use for DINO/BRO).
* BIM and GIS servers;
  + - Open Source BIMServer (developed by TNO and TU Eindhoven).
    - Open Source Deegree Server (built on the standards of the Open Geospatial Consortium (OGC).
    - Various BIM servers from commercial software vendors such as Autodesk (BIM360), Graphisoft, Bentley.
* Enterprise Data Governance (EDG) based on Linked Data;
  + - TopQuadrant TopBraid Enterprise Data Governance (EDG).
    - Ongoing development at TNO ICT (Erik Langius) for ProRail.
* Integrated Development Environments (IDE) for data and functionalities, incl. AI developments;
  + - GitLab and GitHub (open source).
    - MATLAB.
    - Python platforms.
    - Open Source Protégé.
    - TopQuadrant TopBraid Composer.
    - IBM Watson.
* Existing solutions/tools developed by/within TNO BI&M;
  + - TNO InfraTool (developed by Peter Willems).
    - TNO Urban Strategy (currently managed by TNO T&T).
    - TNO Load Map (developed by Pim van den Helm).
    - TNO Siom, DDings as used for green houses.

## Preselection of solutions for SR data ecosystem

The list in section 4.1 has been narrowed down for pre-selecting the possible data ecosystem(s) for SR based on:

1. the general requirements and principles as described in section 2.4;
2. the feedback/experience of the expert users within SR with the currently known/used solutions; and
3. the representativeness of the possible solutions as mapped in the conceptual architecture of SR data ecosystem as previously presented in section 3. The positioning of the various available solutions is shown in the right-side box in the scheme below.

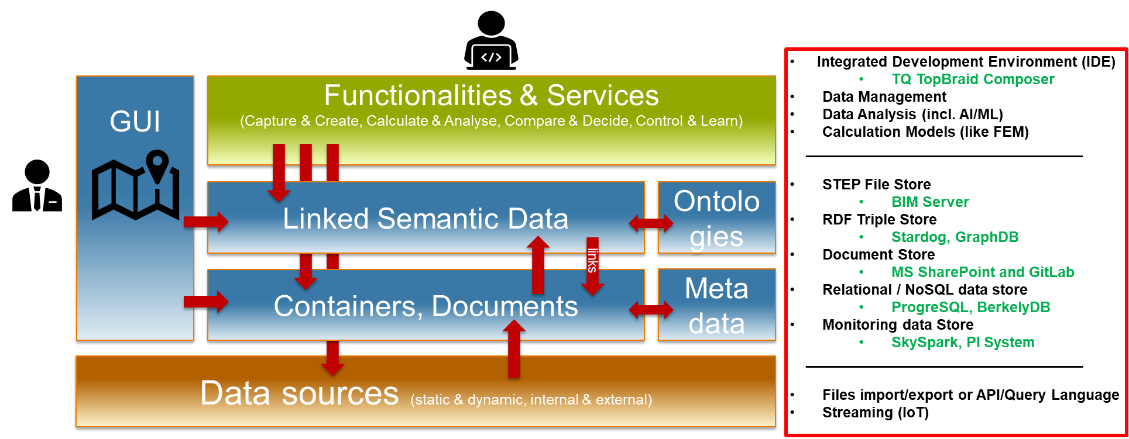


Figure 6: Positioning Data Ecosystem in a Digital Twin concept

*Microsoft SharePoint* for document storage and management is definitely selected since it is a standard solution that has already been decided for the whole TNO.

Several other solutions –as listed below– are preselected based on the aforementioned criteria (a to c). These preselected solutions are prioritised for further investigation and testing / proof-of-concept:

1. *GitLab* for software code and data models and data sets.
2. *OSIsoft PI System* for Digital Twin monitoring data Infra.
3. *SkySpark* for Digital Twin monitoring data Buildings.
4. *Open Source BIMServer* for IFC data.
5. *PostgreSQL* relational database.
6. Semantic database: *Stardog* and/or *GraphDB.*
7. *FIWARE/IDS* (preselected for further investigation considering the interest and ongoing discussions at EU level).

The preselected solutions and the preliminary insights into these solutions up to know are summarised in the following pages.

**GitLab**

GitLab (https://about.gitlab.com/) is a variant of the GitHub software development platform with Git as a basis. Git is widely considered as the default version management system where people can propose changes for software codes in branches that can be committed and merged into the existing codes. GitLab is a more controlled/secure variant of GitHub although both focus on open source software development.

GitLab is an extensive package providing for the complete lifecycle of code development including authentication and authorisation, planning, metric for progress, auditing, issue tracking, testing, etc.

GitLab is meant for managing code but it can also be used for managing data (data sets, data models, etc.). Especially where data documents become too big for Microsoft SharePoint GitLab can be used to collect data sources (like static object data or dynamic monitoring data) and all kinds of derived/enriched data like simulation results.

There are two variants of the GitLab software: a variant that is hosted/serviced by GitLab, and an enterprise edition that can be hosted at an organization (“self-managed” on premises or in the cloud). At TNO we have the Enterprise Edition installed and supported (https://gitlab-dv.tno.nl/users/sign\_in). It is used regularly by experts from SR; and the continued use of it is recommended.

**OSIsoft PI System**

OSIsoft, LLC is a manufacturer of application software for real-time data management, called the PI System. Founded in 1980, OSIsoft is privately held and headquartered in San Leandro, California, USA. PI System is a product of OSIsoft with the ability to collect, analyse, visualize and share large amounts of high-fidelity, time-series data from multiple sources to people and systems across all operations (https://www.osisoft.com/pi-system/).

Working principles:

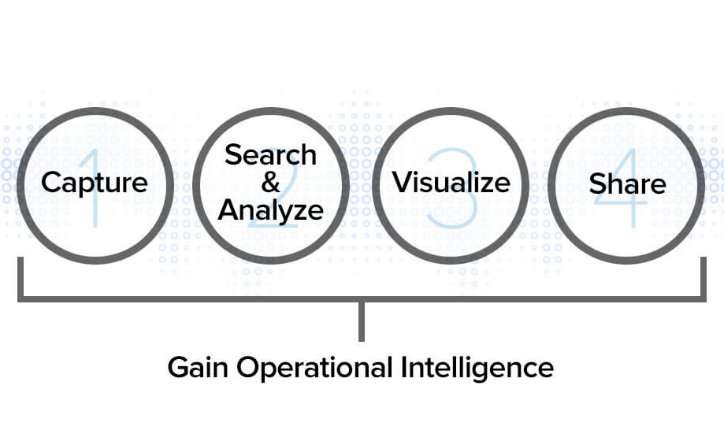


Figure 7: Main working principles of PI System

* Capture: Through an open, scalable architecture and more than 450 off-the-shelf proprietary interfaces to collect high-fidelity time-series data from a diversity of sources –agnostic to system, standard, language, frequency, delivery speed, format or device. It is designed to store millions of data points.
* Search & Analyse: Comparing historical and real-time information to investigate intermittent issues, troubleshoot equipment failures, compare current vs. past production performance and measure new-plant start-ups against existing facilities.
* Visualize: Bringing information to life on any device, prioritizing information, creating context, highlighting salient points, customizing content to different audiences.
* Share: Sharing and collaborating across the enterprise, as well as with partners, vendors and customers; ensuring data security for the right people by enabling fine control of which data is exchanged and who sends and receives it.

Technical overview:

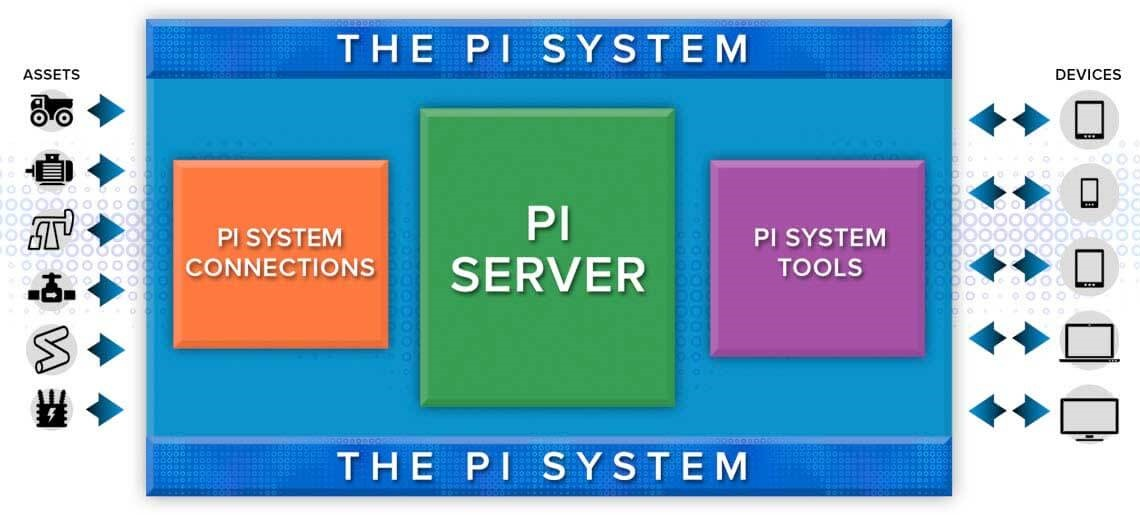


Figure 8: Technical overview of PI System

The current experience of TNO (Joep Paulissen, Henk Miedema et.al.) of using PI System:

* Together with the Bundesanstalt für Materialforschung und -prüfung (BAM) –the scientific and technical Federal institute with responsibility to the Federal Ministry for Economic Affairs and Energy of Germany– the use of the PI System was tested on the measurement data of the “Bridge 705” in Amsterdam. The executed activities are:
* Setting up of link to a BAM WebDav (file server) and a VPN link to the BAM PI System.
* Transformation of three types of measurement data according to the PI input specification:
  + - BREM (micro-strains over time).
    - OSMOS (strains over time).
    - TNO (strains, displacements & accelerations over time).
* Preparing the PI System for this data and the actual import (by BAM).
* Testing the functionality of the system (mainly visualization of the data).
* Findings from this experience:
* Setup was really cumbersome and time consuming. Security is important but a VPN-link from a TNO-network detached system is really unhandy. It worked in the end especially because of the good BAM support.
* The transformation of the data according to PI is non-trivial and time consuming. Enormous amounts of context-sensitive find-and-replace actions are needed for all files. The opposite of “intelligent drag-and-drop” so to say. Learnt from BAM that import was time consuming too.
* PI can handle dynamic (time-series/histories) data only. It seems not a good idea to use the columns to cover the location-based sensor data (many columns).
* Rather simple data visualisation; not many functionalities available other than showing results as they are.

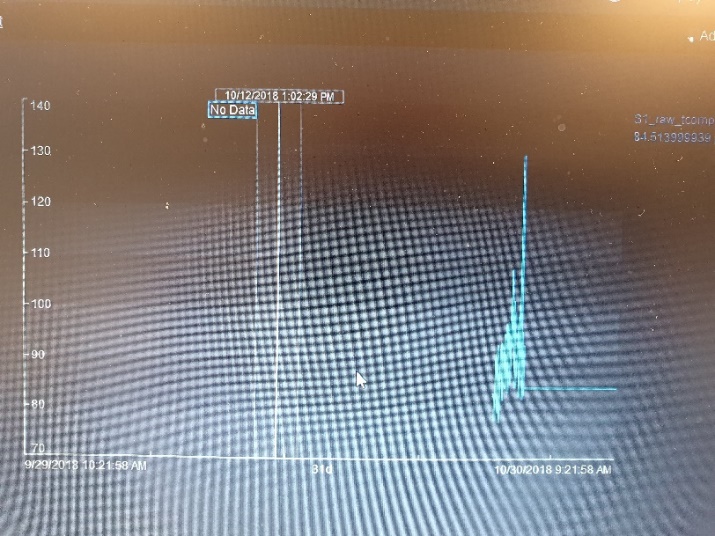


Figure 9: Example of data visualisation in PI System

Recommendations:

* For the short term, like its current usage in the TNO-BAM cooperation, it is recommended to keep on working with the system until additional evaluation results and better information for the alternative solutions become available.
* For the near future, the experiment and testing of the PI System should be continued. The focus should be put on further testing on the actual scale model of the 705 bridge (which is a more relevant usage scenario), in order to make sure that the data has a direct/streaming link with the measurement system. Beware that the differences between the TNO’s and BAM’s measurement systems should be considered to avoid unnecessary effort for manual data transformation and imports.
* For the long term, it is recommended to look for a better solution for measurement data management system, which is available on the market. It would, for instance, be interesting to compare the PI System with more “modern” solutions like the open source InfluxDB (<https://www.influxdata.com/>) as back-end and/or Graphana (<https://grafana.com/>) as front-end.

**SkySpark**

SkyFoundry is a privately held company founded in 2009 and located in Richmond VA, USA. SkyFoundry’s SkySparkTM is a comprehensive software platform for collecting, managing, analyzing and visualizing data from smart devices, sensors and equipment systems (https://skyfoundry.com/product).

SkySpark works with data of all types — whether via a live link to an automation system or smart meter, connection to a SQL database, import of historical data from Excel files, or a web service feed from an utility. The main characteristics are:

* At the heart of SkySpark is the high performance Folio database – specially designed for IoT data.
* SkySpark uses the industry standard Project Haystack for semantic tagging of data. With proper tagging, analytics applications can quickly consume data from equipment systems and interpret patterns in operational data to identify faults, deviations, and trends that can be addressed to improve efficiency and insure proper operation of equipment systems.
* SkySpark uses Axon —an analytic engine that automatically processes rules and algorithms on the data to find Sparks- to automatically generate visualizations, notifications, and reports that show the issue, time of occurrence, frequency, duration, and even cost. It is a way to effectively manage and derive value from the exploding amount of data available from smart and connected devices.
* SkySpark has an Analytic Function Library that includes over 500 built-in analytic functions to create new rules that target the specific needs of the facilities (buildings), equipment, processes and project scope.
* SkySpark includes a rich set of applications to visualize the data and analytic results. All of these intuitive presentations are displayed in standard web browsers using HTML5 technology -no plug-ins required. SkySpark can also output analytic results to 3rd Party applications via its open published APIs.

The current experience of TNO (Bart Driessen, Thijs van der Klauw, Wouter Borsboom et al.) of using SkySpark:

* TNO uses SkySpark for Digital Twin research on building’s energy efficiency and indoor climate, with a field lab of TNO’s office on Waalsdorperweg in The Hague.
* TNO has a limited license for SkySpark that allows it to connect the platform with data points from sensors to measure temperature, people’s movements inside the building, humidity, CO2, etc.
* The acquired data is stored, processed and visualised on SkySpark.

The findings from this experience can be summarised as follows:

* SkySpark can be interfaced with various programming languages with which TNO experts are familiar, such as Phyton and Java. Other flexibility aspects are also possible, for instance: data processing in Phyton and data visualisation in SkySpark.
* Inside SkySpark there is a development environment that can be dedicated to create specific tools, for instance the connection with the control system / actuators.
* SkySpark supports open standard Project Haystack with its growing use in the building sector to recognise/register IoT devices.
* SkySpark has a library of tools/apps and standard connections with weather data, etc., which can save time for configuring the needed software. This is also a positive aspect regarding user-friendliness. However, not all needed tools are there, so effort for tool development depending on the specific purposes is still needed.
* SkySpark has a collection of case studies, white papers and an active user community for discussions and practical questions.

Recommendations:

* Further and deeper exploration of SkySpark is worth it since this platform is gaining popularity especially in the domain of smart buildings and building’s energy management. It also has the advantage of the possible connection with other IoT platforms known in the other industrial sector.
* For SR research purposes in energy performance of buildings, SkySpark is suitable, yet further investigation is needed on the possible solutions to the current limitations of SkySpark, i.e. it cannot handle IFC BIM data, and it is currently not known / not used in the civil infrastructure sector.
* A detailed analysis can be found in the master thesis by Mathias Middernacht from KU Leuven (2018) titled “Prestatie visualisatie van een smart building”.

**BIMServer**

Among the many data sets we want to be able to handle, ISO STEP Physical Files structured according to the open BIM IFC standard are important. TNO has together with the TU Eindhoven contributed to the java-based open source BIMServer (bimserver.org). This server can import and store IFC data in an underlying key-value Berkeley Database (BDB). Key functionality is the visualisation of an IFC BIM model.

In the BIMServer, data can be accessed via IFC STEP file import/export or vmore directly via SOAP. Also other derived serialisations (XML, linked data) are supported to some extent.

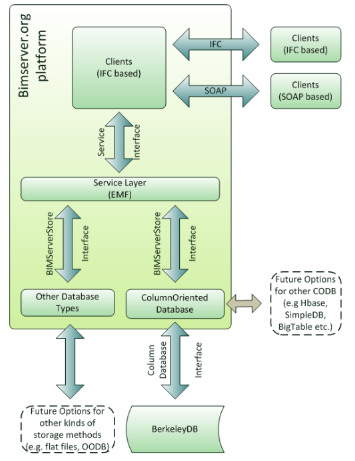


Figure 10: Schematic overview of the OS BIMServer

Currently, the BIMServer is strongly related with the further developments and usage of “BIM bots” by TNO, microservices that work on data with specific BIM interfaces. It seems logical to start with the BIMServer, but we should have an open eye for external alternatives since the BIMServer still requires significant TNO’s investments to make the historic patch-wise developments into a robust and stable system. Besides the open source BIMServer many alternatives exist, including:

https://ifcwebserver.org/ and https://bim-plattform.com/en/bimq/.

**PostgreSQL**

Many databases used in practice are relational (compared to the more modern NoSQL or semantic databases). PostgreSQL (https://www.postgresql.org/) is the open source (object-)relational database management system based on the Structured Query language (SQL) standard ISO/IEC 9075:2016. It originated from the University of California in Berkeley and has a 30 year active development history.

It runs on all major operating systems and is very popular in GIS context because of its popular PostGIS extension.

PostgreSQL is feature-rich covering various data management aspects:

Data Types:

* Primitives: Integer, Numeric, String, Boolean
* Structured: Date/Time, Array, Range, UUID
* Document: JSON/JSONB, XML, Key-value (Hstore)
* Geometry: Point, Line, Circle, Polygon
* Customizations: Composite, Custom Types

Data Integrity:

* UNIQUE, NOT NULL
* Primary Keys
* Foreign Keys
* Exclusion Constraints
* Explicit Locks, Advisory Locks

Concurrency, Performance:

* Indexing: B-tree, Multicolumn, Expressions, Partial
* Advanced Indexing: GiST, SP-Gist, KNN Gist, GIN, BRIN, Covering indexes, Bloom filters
* Sophisticated query planner / optimizer, index-only scans, multicolumn statistics
* Transactions, Nested Transactions (via save points)
* Multi-Version concurrency Control (MVCC)
* Parallelization of read queries and building B-tree indexes
* Table partitioning
* All transaction isolation levels defined in the SQL standard, including Serializable
* Just-in-time (JIT) compilation of expressions

Reliability, Disaster Recovery:

* Write-ahead Logging (WAL)
* Replication: Asynchronous, Synchronous, Logical
* Point-in-time-recovery (PITR), active standbys
* Tablespaces

Security:

* Authentication: GSSAPI, SSPI, LDAP, SCRAM-SHA-256, Certificate, and more
* Robust access-control system
* Column and row-level security
* Multi-factor authentication with certificates and an additional method

Extensibility:

* Stored functions and procedures
* Procedural Languages: PL/PGSQL, Perl, Python (and many more)
* SQL/JSON path expressions
* Foreign data wrappers: connect to other databases or streams with a standard SQL interface
* Customizable storage interface for tables
* Many extensions that provide additional functionality, including PostGIS

Internationalisation, Text Search:

* Support for international character sets, e.g. through ICU collations
* Case-insensitive and accent-insensitive collations
* Full-text search

Although it is not fully capable of handling real-time Big Data, PostgreSQL is still relevant for SR. It should be kept for specific programming work rather than for a broad usage.

**Semantic database: *Stardog* and *GraphDB***

Linked Data / Semantic Web technology is commonly considered as the future for data modelling and linking. Therefore preselecting and testing of two currently popular solutions in the market are recommended, especially for: Stardog (also used in our European BIM4Ren project) and GraphDB (chosen by Rioned as GWSW server (sewer data) after analysis of major semantic platforms on functionality, performance and pricing).

*OntoText GraphDB (*http://graphdb.ontotext.com/) is a java-based (RDF4J) triple store that comes into various editions including a free edition. All versions support full SPARQL1.1 query support. GraphDB provides extensive reasoning capabilities based on OWL-RL and OWL-QL variants. A special functionality is the built-in support for GeoSPARQL. The more advanced editions support higher scalability via parallelisation and load-balancing over clusters. There is only experimental SHACL support (closed world alternative for data verification to OWL).

Stardog (https://www.stardog.com/) is a similar solution that offers more advanced functionalities especially in their workbench called Stardog Studio involving data and query visualisation.

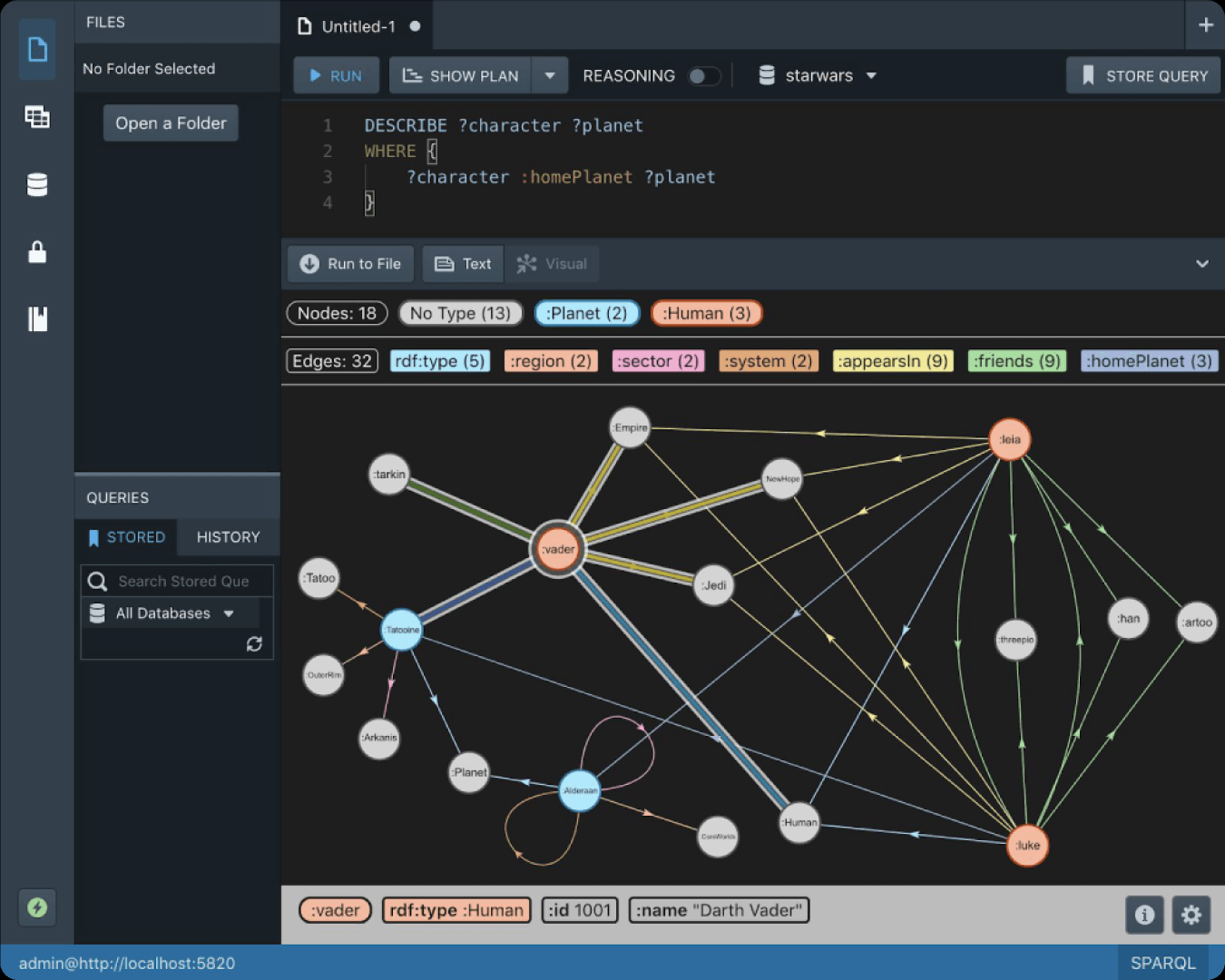


Figure 11: Data visualisation in Stardog

Stardog and GraphDB support the strict W3C standards (RDF, RDFS, OWL, SPARQL. etc.). Both systems are recommended for use in SR in the near future along with the development of needed knowledge and skills of SR experts. An advantage of using these systems regarding the data sets is: the direct reusability of the data sets from one system to test the other. Additionally, GraphQL is also a more simple front-end where SPARQL querying is supported.

**FIWARE / IDS**

FIWARE is an open source initiative defining a universal set of standards for context data management which facilitate the development of Smart Solutions for different domains such as Smart Cities, Smart Industry, Smart Agrifood, and Smart Energy. In any smart solution there is a need to gather and manage context information, processing that information and informing external actors, enabling them to actuate and therefore alter or enrich the current context.

FIWARE is focused on industry reference architecture, which is Compliant with existing industry architectures such as the Reference Architecture Model Industry 4.0, the Industrial Data Space Reference Architecture or the Industrial Internet Consortium Reference Architecture. The schematic representation of FIWARE

Ecosystem can be found in the following figure

(Ref: https://www.fiware.org/community/smart-industry/).

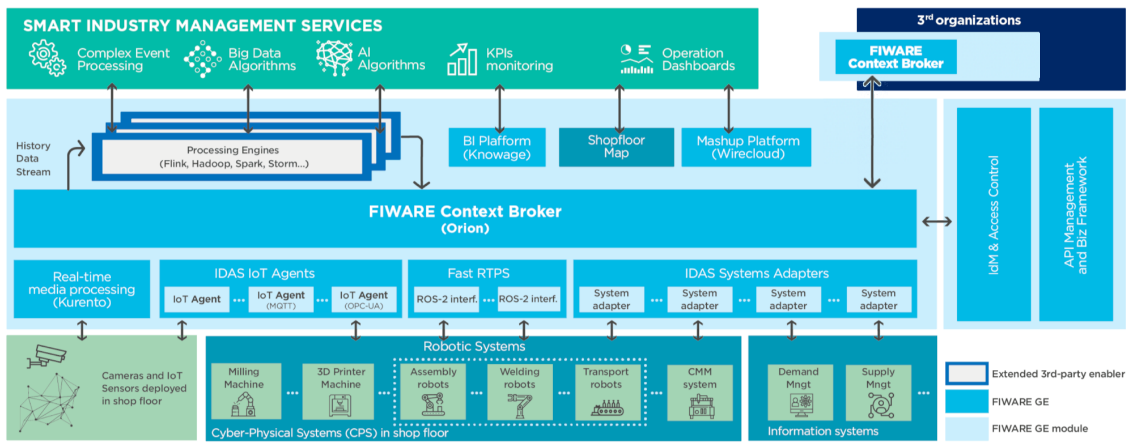


Figure 12: Schematic representation of FIWARE ecosystem

The FIWARE Context Broker component is the core component of any “Powered by FIWARE” platform. It enables the system to perform updates and access to the current state of context. The Context Broker in turn is surrounded by a suite of additional platform components, which may be supplying context data (from diverse sources such as a CRM system, social networks, mobile apps or IoT sensors for example), supporting processing, analysis and visualization of data or bringing support to data access control, publication or monetization

(Ref: https://www.fiware.org/developers/).

FIWARE and the International Data Spaces Association (IDSA) are working on the open source implementation of the Industrial Data Space (IDS). IDS addresses a key topic in the evolution of Industry 4.0: How can companies and institutions build a space where data is shared in a decentralized manner so that each organization can use available data to improve their processes, as well as govern and monetize data exported to third parties. The International Data Spaces Association (IDSA) is creating a reference architecture to implement secure and trustworthy data exchanges. Data providers hereby keep control over the usage of their data, also referred to as “data sovereignty” (Ref: https://www.internationaldataspaces.org/wp-content/uploads/dlm\_uploads/2018/08/IDSA-Infografik-EN.pdf).

FIWARE and IDS were both named by the European Union (EU) with regards to the progress of the Digitising European Industry (DEI) initiative. TNO is a member of the IDSA (experts a.o. Frans van Etten and Matthijs Punter).

However, this revolution will also affect business and industry bringing the IoT to the production processes in what is called Industry 4.0. Sensor-enabled manufacturing equipment will allow real time communication, smart diagnosis and autonomous decision making. In this scope, the

Industrial Data Spaces (IDS) is a Reference Architecture model that aims to provide a common frame for designing and deploying Industry IoT infrastructures (i.e. sensor-enabled manufacturing equipment will allow real time communication, smart diagnosis and autonomous decision making). Such a Reference Architecture can be implemented based on FIWARE open source software components (Generic Enablers). FIWARE-based IDS implementation fits the requirements of the IDS Reference Architecture providing open source software suitable to any Industry 4.0 environment (Ref: Alonso, A., Pozo, A., Cantera, J.M., Vega, F. de la, Hierro, J.J. (2018), Industrial Data Space Architecture Implementation Using FIWARE, in: Sensors (Basel). 2018 Jul; 18(7): 2226).

Regarding the search for a data ecosystem for SR purpose, FIWARE and IDS are not a platform solution to be used as they are rather a reference architecture which will be useful for the development of an ecosystem, platform or software. Since SR neither intends nor it is able to develop a new ecosystem or platform (and therefore, it is looking for available solutions), FIWARE and IDS are not immediately relevant for SR at present. Nevertheless, gaining the knowledge and following the developments –in collaboration with experts from TNO ICT– are surely valuable as these initiatives are endorsed by the European Commission for a possible wider use in various domains, including smart building and construction.

# Conclusions and recommendations

## Conclusions

The goal of this study is to propose a solution for a data ecosystem that facilitates and supports research activities by TNO’s expertise group Structural Reliability (SR).

A short-list of relevant solutions is presented in section 4.1 in this report and a long-list is included in Appendix 1. Based on the general requirements and fundamental principles summarised in section 2.4 as well as the conceptual IT architecture according to SR’s needs presented in chapter 3, several preferred solutions are selected. These solutions have been described and assessed in section 4.2. The partial conclusion for each possible solution is included in the same section.

The main conclusion that can be drawn from this study is that there is no single existing data ecosystem or platform solution that fully covers all the needs of SR within the given boundary conditions (incl. cost considerations that will be assessed in the follow-up research). Therefore, the feasible solution is a combined use of (parts of) several ecosystems/platforms.

Next to the ‘TNO standard’ ecosystems that will be continued to be used (such as Microsoft SharePoint for document management), the positioning and coverage of the selected solutions in a combined ecosystem can be mapped on the conceptual SR IT architecture as shown in the following figure.

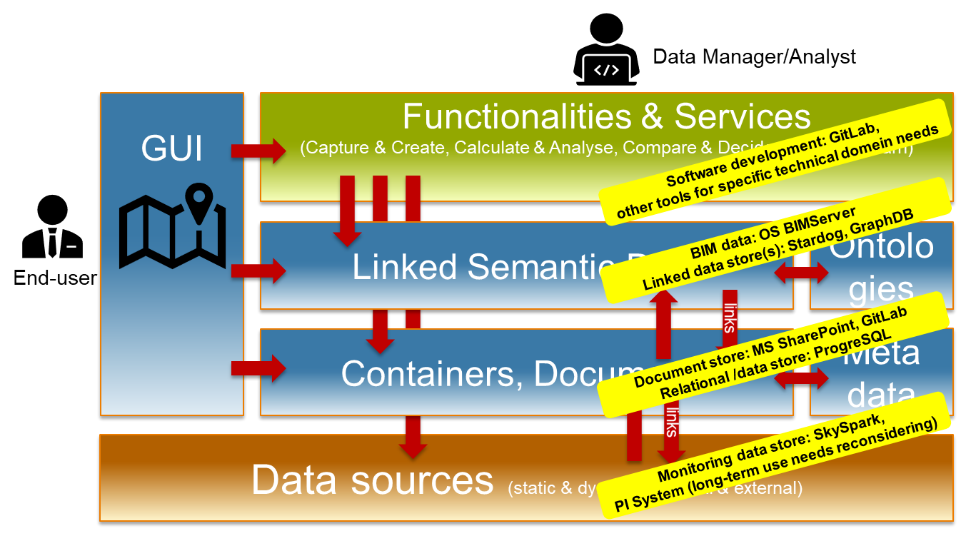


Figure 13: Mapping of the combined solutions that cover the SR conceptual IT architecture

## Recommendations

As follow-up of this study, three main recommendations are given:

* The selected combined solutions should be subject to a further / in-depth assessment based on the more detailed criteria listed in section 2.5 by applying it in SR’s research projects.
* The integration approach over the selected combined solutions should be further developed. Such an integration approach should cover two aspects:
  1. Technical integration between two or more selected ecosystems within the combined solutions. Some technical developments currently take place within the on-going EU research projects, for instance in BIM4Ren where a technical integration solution between BIM (IFC in STEP technology) and Linked Data (ontologies in W3C technology) is developed. However, for a comprehensive purpose for SR, such integration effort should be harmonised and should not be only dependent on the research plans of the individual (EU) projects.
  2. User / process integration through which the expert users are guided when utilizing the different ecosystems within the combined solutions. Such user / process integration can be promoted both by setting up a coherent guideline as well as by facilitating discussions between the key users of the different ecosystems. Through such an integration, an expert user will be encouraged and assisted to consider the re-use and integration potential of the data by the other ecosystems beyond the one that he/she uses.
* The combined solutions should be scaled up beyond SR. The short-term priority could be the use of the combined solutions by multiple departments within the BI&M unit, for instance in the collaboration within the on-going ERP Digital Twin - Structural Integrity. A further upscaling should subsequently be considered to involve TNO’s clients / external collaboration partners in the various research projects, especially related to Digital Twin.

# Signature

Delft, 17 December 2019 TNO

Dr. Ir. H.M. Böhms

Lead Author

Dr. P.C. Rasker

Research Manager Structural Reliability

# Appendix 1: Benchmarking Existing Smart Cities Platforms

By Yacine Rezgui, Sylvain Kubicki and Alain Zarli (2019)

References:

* Kolozali S, Bermudez-Edo M, Puschmann D et al. A knowledge-based approach for real-time IoT data stream annotation and processing. 2014 IEEE International Conference on Internet of Things (iThings), and Green Computing and Communications (GreenCom), IEEE and Cyber, Physical and Social Computing (CPSCom), Tapei, Taiwan, 1–3 March 2014. New York, IEEE, 2014, pp 215–222.
* Almanac. D6.1: A scalable data management architecture for smart city environments. Torino, Italy, Almanac Consortium, 2014, 43pp.
* RERUM. D2.3: System architecture. RERUM FP7-ICT-609094. RERUM Consortium, Heidelberg, Germany, 2014.
* Lea R and Blackstock M. City Hub: a cloud-based IoT platform for smart cities. 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, Singapore, 15–18 December 2014. New York, IEEE, 2015, pp 799–804.
* Kolozali S, Puschmann D, Karapntelakis A et al. D3.1: Semantic data stream annotation for automated framework. Guildford, Surrey, CityPulse, 2014, 38pp.

A wide range of IoT platforms have been developed by the research community. However, limited examples were found which integrate closely with semantics at building and district level such as Building Information Models (standardized through IFC format) and City Information Model (standardized e.g. through CityGML). These projects emphasise aspects such as scalable stream processing, including semantic tagging, but only regarding the data and sensors themselves, rather than their physical built environment context. The ALMANAC project did include some physical concepts, but only to a basic level, rather than the detailed approach of BIM.

Figure A1 illustrates the ideal Semantic Web of Things architecture that extends the traditional IoT stack.

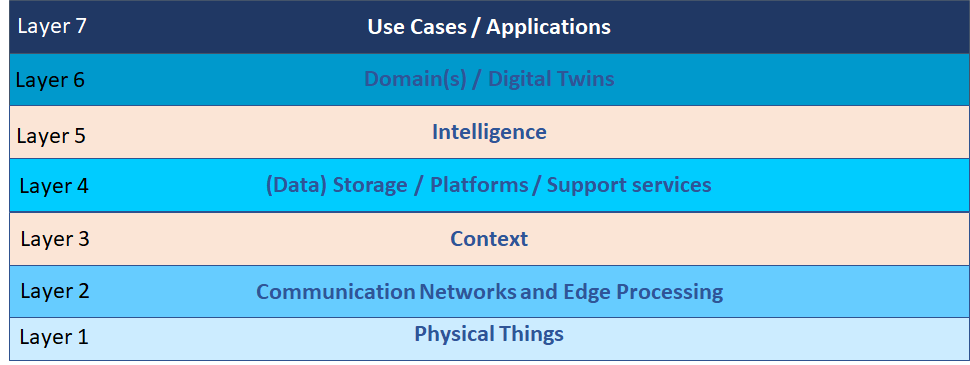


Fig. A1. Generic IoT for Smart City Architecture

The above layers are briefly described below:

* **Layer 1: Physical Things**

This layer refers to a wide range of connected objects, such as sensor, actuators, user data through smart phones, etc.

* **Layer 2: Communication Networks and Edge Processing**

This layer deals with the heterogeneity of the current IoT protocols and as such provides an interface to upper layers while abstracting the complexity of the underpinning formats. In some instances, this layer is augmented with Edge Processing which helps alleviate the processing burden on the upper layers and as such result in a decentralized and more scalable architecture.

* **Layer 3: Context**

Context refers to a partial or complete description of the domain or domains (in a complex system of systems scenario, such as the one required for smart cities). The context can therefore range anything from application/use case driven, using simple data structures, that provide a logical model of the phenomenon at hand, to more holistic conceptualizations that draw on the IFC/CityGML and other existing models. Table A1 illustrates a subset of relevant models providing context for smart cities.

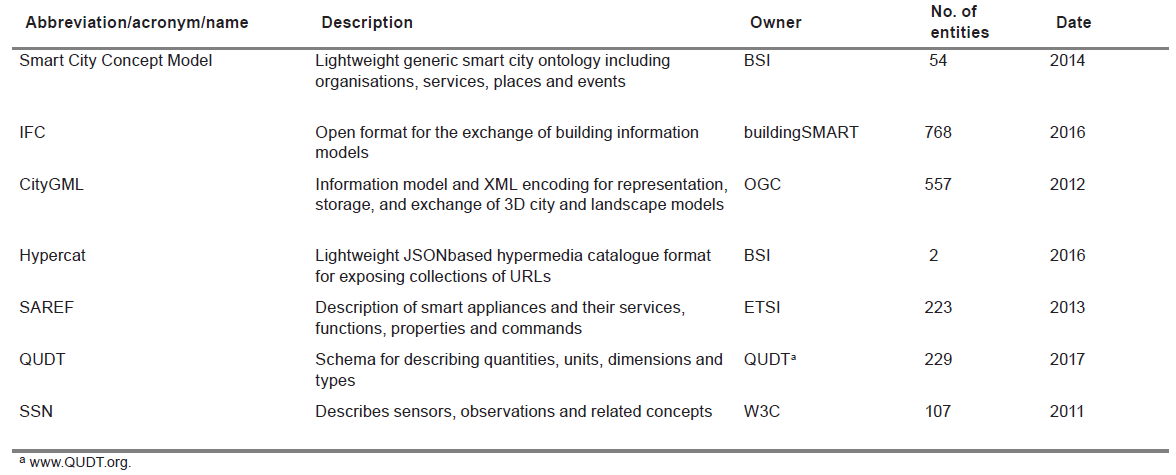


Table A1. Some existing models for IoT Platforms

* **Layer 4: Storage**

The data storage refers to collection, storage and processing of all raw sensor data. This also involves historic data analysis and end users access to big data through adapted APIs. Sensed raw data can be stored on a local (proprietary) solution, in cloud-based services, or use a hybrid system.

* **Layer 5: Intelligence**

Intelligence involves the use of Machine Learning algorithms. It is often relegated to Application level (Level 7). The use of intelligence as part of an IoT middleware is uncommon.

* **Layer 6: Domain / Digital Twins**

This layer represents the domain in which the data is being collected. This can range from a discrete sub-domain, to a given domain, or even involve several domains as is the case in Smart Cities applications, such Energy, Water, Mobility, etc.

* **Layer 7: Use Case Application**

A given domain may involve several use cases, such as (a) an energy demand/response scenario and (b) energy consuming devices scheduling scenario.

Based on the above layers, Table A2 illustrates key features of each IoT for Smart City platform identified in this paper.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Smart Cities Platform** | **Description** | **Layer 1** | **Layer 2** | **Layer 3** | **Layer 4** | **Layer 5** | **Layer 6** | **Layer 7** |
| CityPulse | Combines a knowledge-based approach with reliability testing to provide a platform based  smart city IoT applications. | v | v |  | v |  |  | v |
| ALMANAC | IoT platform which collects, aggregates and analyses real-time or near real-time data from  heterogeneous sensors and actuators to support Smart City processes. | v | v |  | v | v |  | v |
| RERUM | Smart city IoT platform with an emphasis on security and coping with heterogeneity | v | v |  | v | v |  | v |
| VITAL | Heterogeneous IoT system and service integration project. | v | v |  | v | v |  | v |
| CLOUT | Cloud Computing project aiming to bridge IoT with ‘Internet of People’, emphasis on  effective integration. | v | v |  | v | v |  | v |
| SMARTIE | Security, privacy and trust project for consumer IoT data. | v | v |  | v | v |  | v |
| FIESTA | IoT platform for heterogeneous IoT technology integration. Emphasis on a semantics based  solution and providing more than just software outputs. | v | v |  | v | v |  | v |
| FIWARE | Open Source IoT platform for Smart Solutions | v | v | Limited | v |  | v | v |
| IMEC  (TNO) | Internet of Things for cities | v | v |  | v |  | v | v |
| IBM Watson | Data analytics platform | v | v | Limited | v | v | v | v |
| SIEMENS | IoT augmented with BIM (Bentley) | v | v | Limited | v | v | v | v |
| Johnson  Systems | IoT Platform | v | v |  | v | Limited | v | v |
| Azure IoT | IoT service that powers comprehensive virtual representations of physical environments and associated devices, sensors, and people | v | v | Limited | v | Limited | v | v |
| Priva ECO | A cloud-based service that can be implemented as an add-on to existing building automation systems from any vendor | v | v |  | v | Limited | v | v |
| CUSP | IoT underpinned by Semantics and powered by AI | v | v | v | v | v | v | v |

Table A2. Key features of existing IoT Smart Cities Platforms.

1. <https://www.force11.org/group/fairgroup/fairprinciples> [↑](#footnote-ref-2)