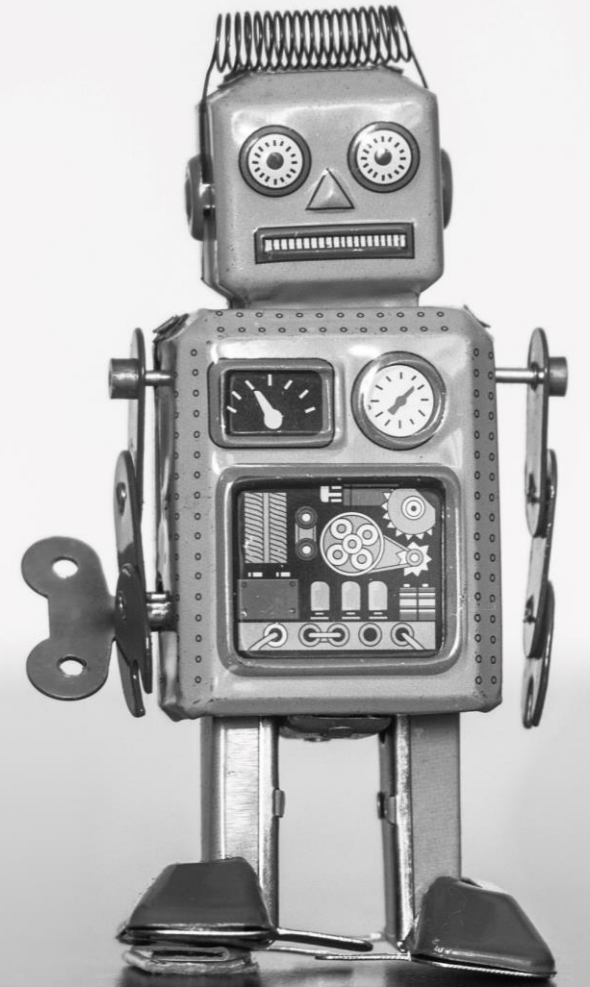




DATA GOVERNANCE IN THE AGE OF AI

How AI impacts your data landscape

Jan Voskuil | Taxonic
Kenniskring | 3 juni 2024
Wageningen





TAXONIC

WHO WE ARE

Founded in **2012**, we specialize in knowledge engineering and practical application of knowledge graphs

Our mission: leverage **model-driven** architectures

Reseller for **TopQuadrant's** TopBraid suite

Pega Systems partner

Currently **37 consultants** on the payroll



Linked Data




Dynamic Case Management




ICT Dienstverlening



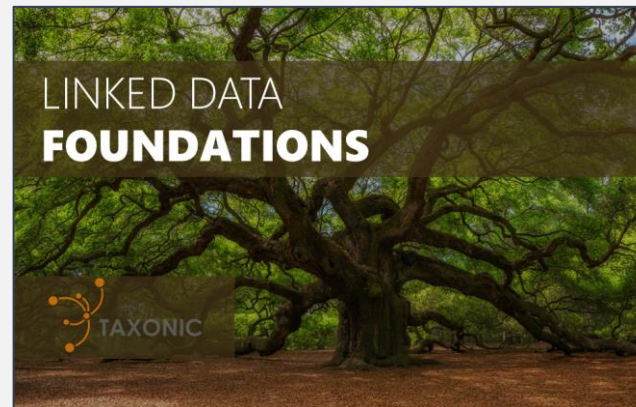
PART OF OUR MISSION: **SHARING KNOWLEDGE**




LINKED DATA
ESSENTIALS



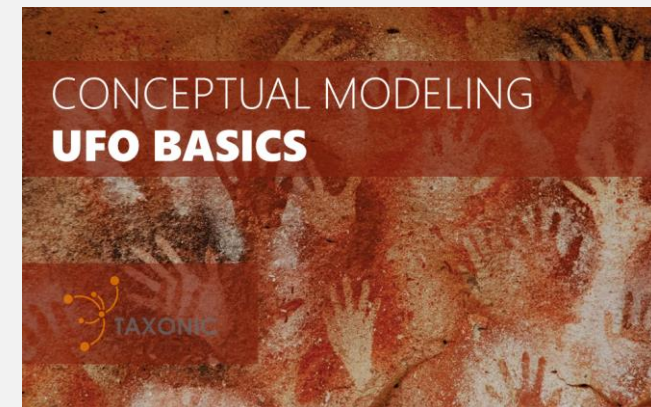
E-LEARNING
OPTIONAL WORKSHOP AVAILABLE
8 hours




LINKED DATA
FOUNDATIONS



E-LEARNING
OPTIONAL WORKSHOP AVAILABLE
24 hours



CONCEPTUAL MODELING
UFO BASICS



E-LEARNING
OPTIONAL WORKSHOP AVAILABLE
8 hours

www.taxonicacademy.com


1

Chapter

1. The problem with AI
2. How to improve AI
3. Why AI matters



Oude technologie hindert banken bij gebruik AI

 Rutger Betlem

In de boardrooms van de grootbanken kunnen ze niet om AI heen. Experts beschrijven het als haarlemmerolie, een wondermiddel dat banken beter, sneller en goedkoper maakt. Van het oplossen van bureaucratie en beter portfoliomanagement tot het verhogen van de efficiëntie. Maar gaat dat ook lukken? Niet iedereen is overtuigd.



AI is nieuw, de systemen waar grootbanken op draaien zijn oud. Dat gaat moeilijk samen. Illustratie: iStock/FD Studio

1

AI IS NOT READY

2

DATA IS NOT READY

3

THE ORGANIZATION IS NOT READY

BAREND MONS: "STOP DATA SHARING"

(shared keynote with [ICBO](#))

The rapid developments in the field of machine learning have also brought along some existential challenges, which are in essence all related to the broad concept of 'trust'. Aspects of this broad concept include trust in the output of any ML proces (and the prevention of black boxes, hallucinations and so forth). The very trust in science is at stake, especially now that paper mills come up that also aggravate the perverse reward systems in current research environments, which are stuck in 20th (in fact 17th) century scholarly communication. The other side of the same coin is that ML, if nor properly controlled will also break through security and privacy barriers and violate GDPR and other Ethical, Legal and Societal barriers, including equitability. In addition, the 'existence' of data somewhere by no means implies its actual Reusability. This includes the by now well established four elements of the FAIR principles: Much data is not even **F**indable, if found, not **A**ccessible under well defined conditions, and if accessed not **I**nteroperable (understandable by third parties and machines) and this results in the vast majority of data and information not being **R**eusable without violation of copyrights, privacy regulations or the basic conceptual models that implicitly or explicitly underpin the query or the deep learning algorithm. This keynote will address how 'data visiting' as opposed to classical 'data sharing', which carries the connotation of data downloads, transport and loosing control, mitigates most, if not all, the unwanted side effects of classical 'data sharing'. For federated data visiting, the data should be FAIR in an additional sense or perspective, they should be '**Federated, AI-Ready**', so that visiting algorithms can answer questions related to Access Control, Consent, Format, and can read rich (FAIR) metadata about the data itself to determine whether they are 'fit for purpose' and machine actionable (i.e. FAIR digital Objects, or Machine Actionable Units). The 'fitness for purpose' concept goes way beyond (but includes) information about methods, quality, error bars etc. The 'immutable logging' of all operation of visiting algorithms is crucial, especially when self learning algorithms in 'swarm learning' are being used. Enough to keep us busy for a while.



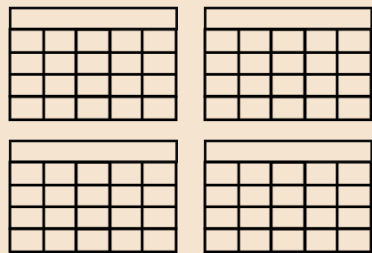
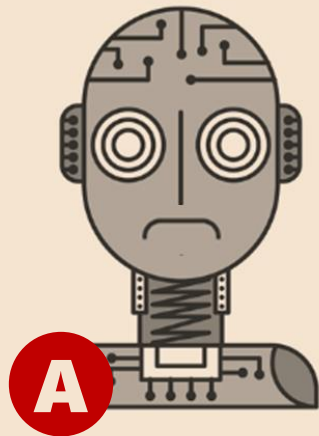
F A I R

Findable, Accessible,
Interoperable, Reusable

F A I R

Federated, AI Ready


IMPROVING AI WITH KNOWLEDGE GRAPHS




- > RDMS → knowledge graph
- > Take two identical AIs, A and B
- > Train A on the RDMS
- > Train B on the KG
- > Measure performance

KGs improve AI's performance

KGs lead to Explainable AI



Information managers and architects
Who are not busy with AI
Are working on yesterday's
architecture

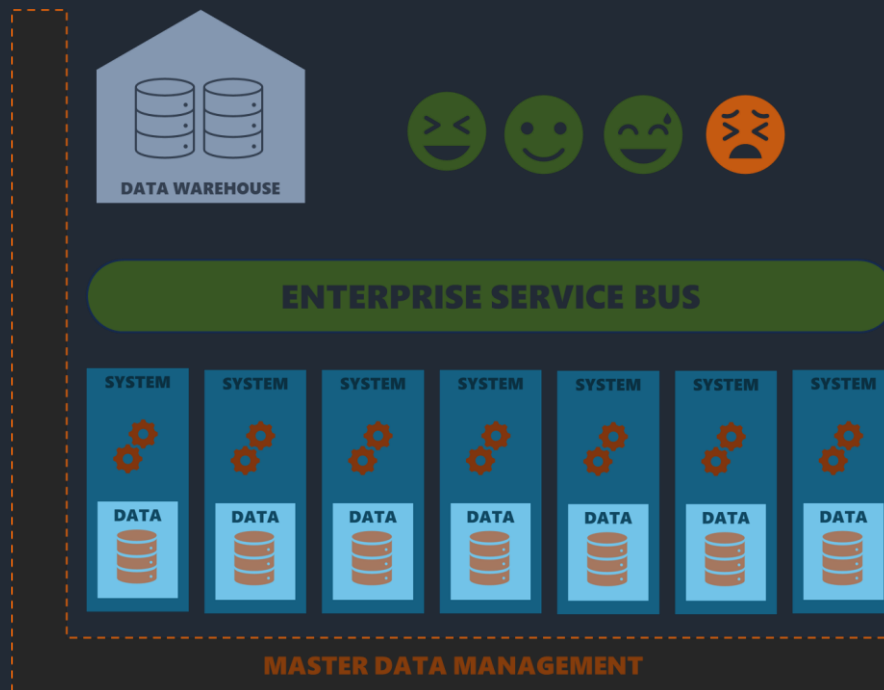


2

Chapter

1. The problem with data
2. How to improve data
3. Why this is necessary

DATA CENTRIC THE NEW ARCHITECTURE



Problem: Process Centricity

- > Silo's are an effect of the problem
- > ESB, MDM, DW do not solve it
- > Separate index card systems

Solution: Data Centricity

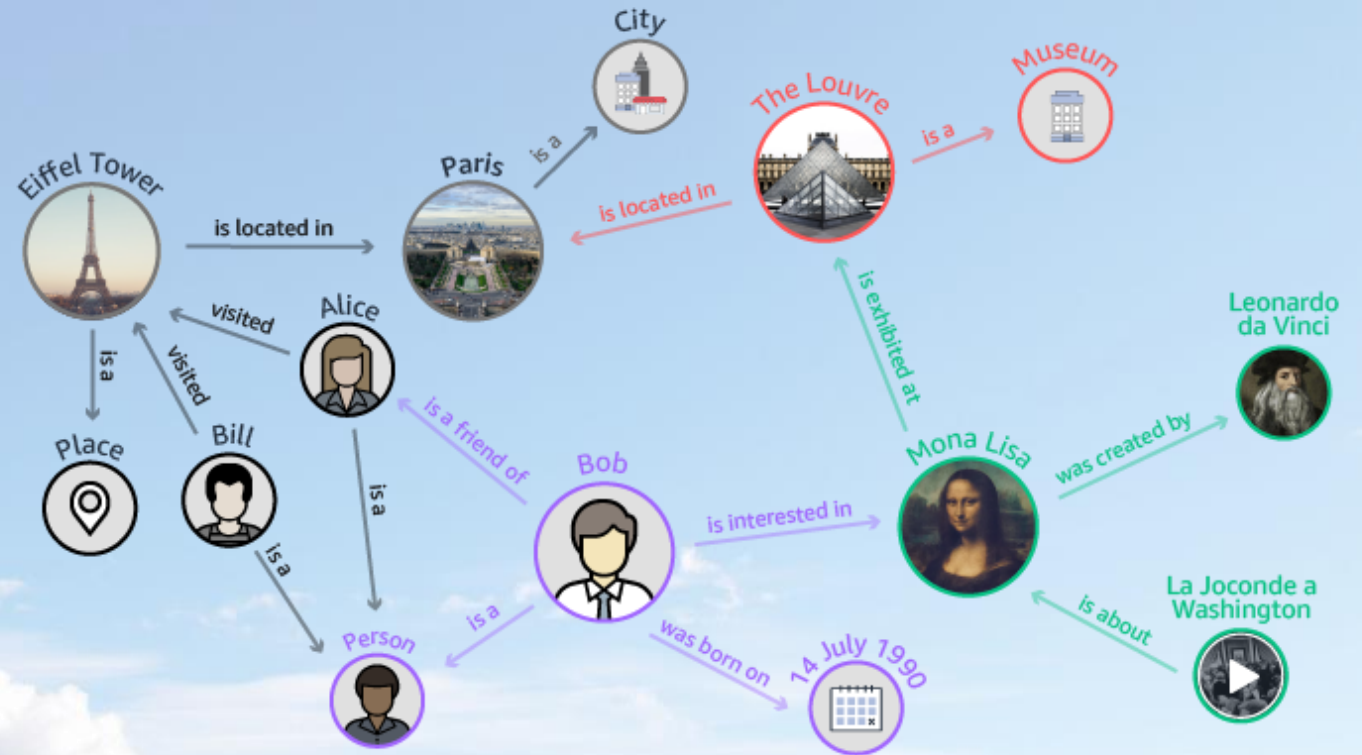


Dave McComb
The Data Centric Manifesto

DATA KNOWLEDGE GRAPH

RDF Knowledge Graph (1999)

- > All identifiers are http URIs
- > All facts are represented as triples



ANATOMY OF A KNOWLEDGE GRAPH

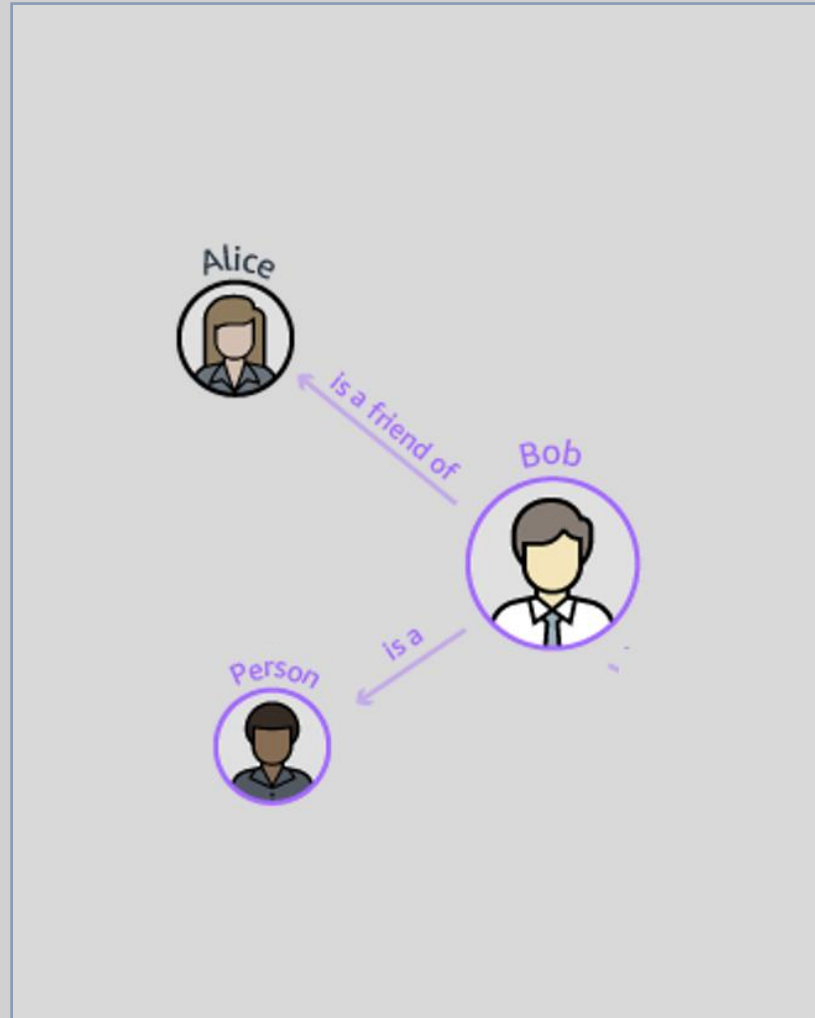
TOWARDS A DATA CENTRIC LANDSCAPE

- > Read **from** and write **to** the KG
- > Applications use the same model
- > App-internal duplication is OK
- > Not **all** data is necessarily in the KG

Subject	Predicate	Object
Paris	is a	City
Eiffel Tower	is located in	Paris
Alice	visited	Eiffel Tower
Alice	is a	Person
Bill	visited	Eiffel Tower
Bill	is a	Person
Eiffel Tower	is a	Place
Bob	is a	Person
Bob	is a friend of	Alice
Bob	is interested in	Mona Lisa
Mona Lisa	is exhibited in	The Louvre
The Louvre	is located in	Paris
...

KNOWLEDGE GRAPH

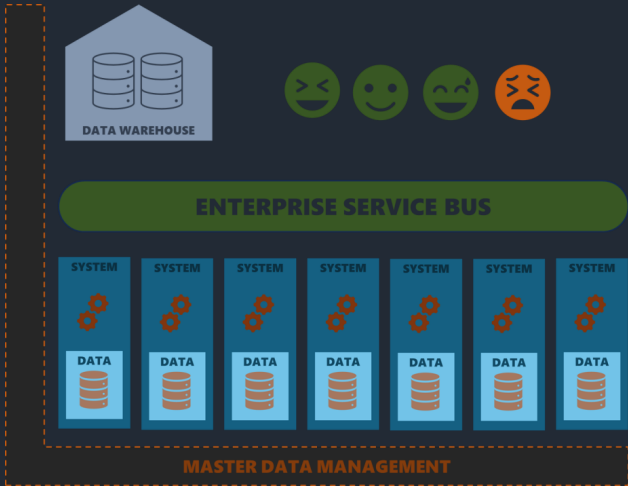
MODELS + DATA



Subject	Predicate	Object
Bob	is-a-friend-of	Alice
Bob	is-a	Person

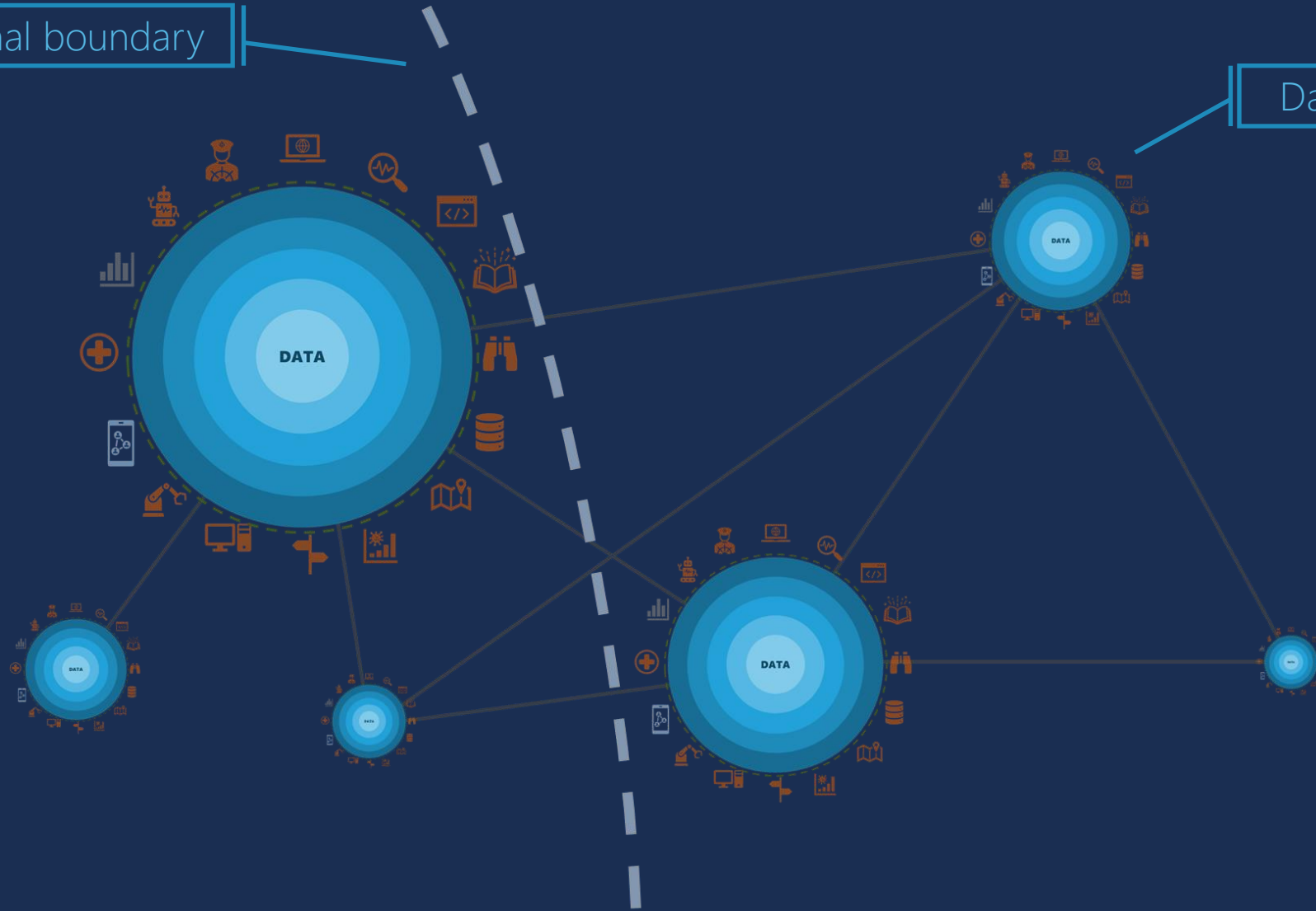
Data is represented as **data**
Model is represented as **data**

THE LONG ROAD TOWARDS DATA CENTRISM



Organizational boundary

Data as a Service



ANATOMY OF A KNOWLEDGE GRAPH

ONTOLOGY

Data model: what there is

RULES

Age = today minus birth date
Customer has telephone number

REFERENCE DATA

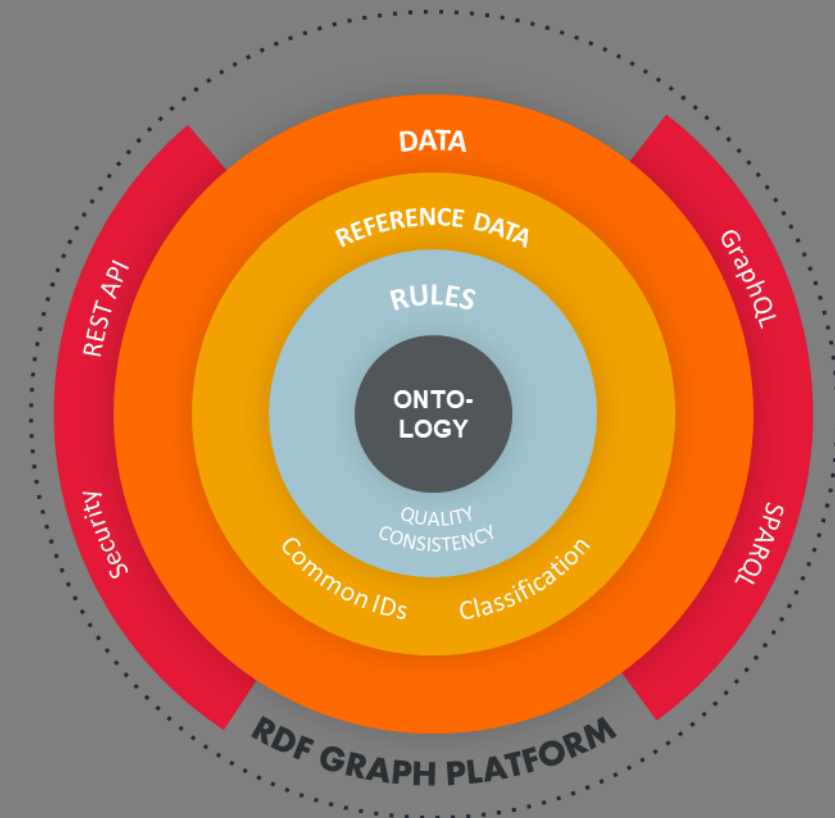
Municipality of Amsterdam
gemeente : gm0363

DATA

Modularized in subgraphs

SERVICES

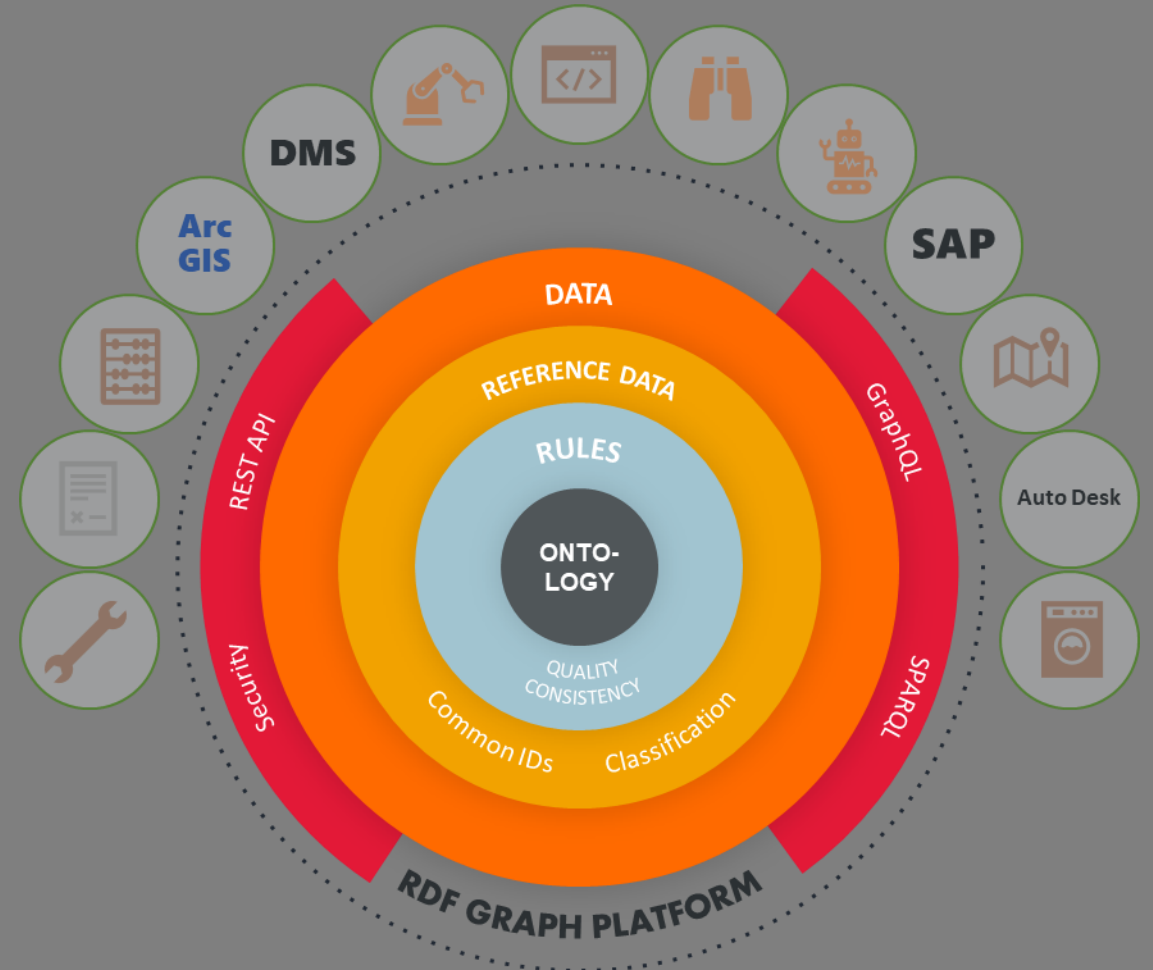
Security, APIs, Querying



ANATOMY OF A KNOWLEDGE GRAPH

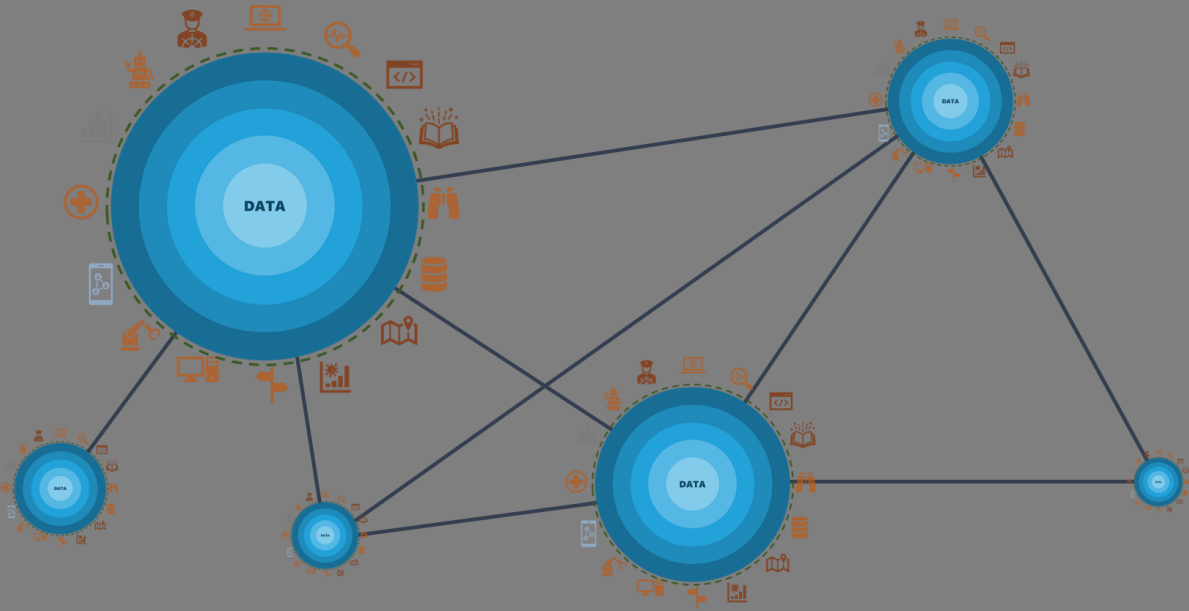
TOWARDS A DATA CENTRIC LANDSCAPE

- > Read **from** and write **to** the KG
- > Applications use the same model
- > App-internal duplication is OK
- > Not **all** data is necessarily in the KG



RELEVANCE

SILOS DO NOT WORK



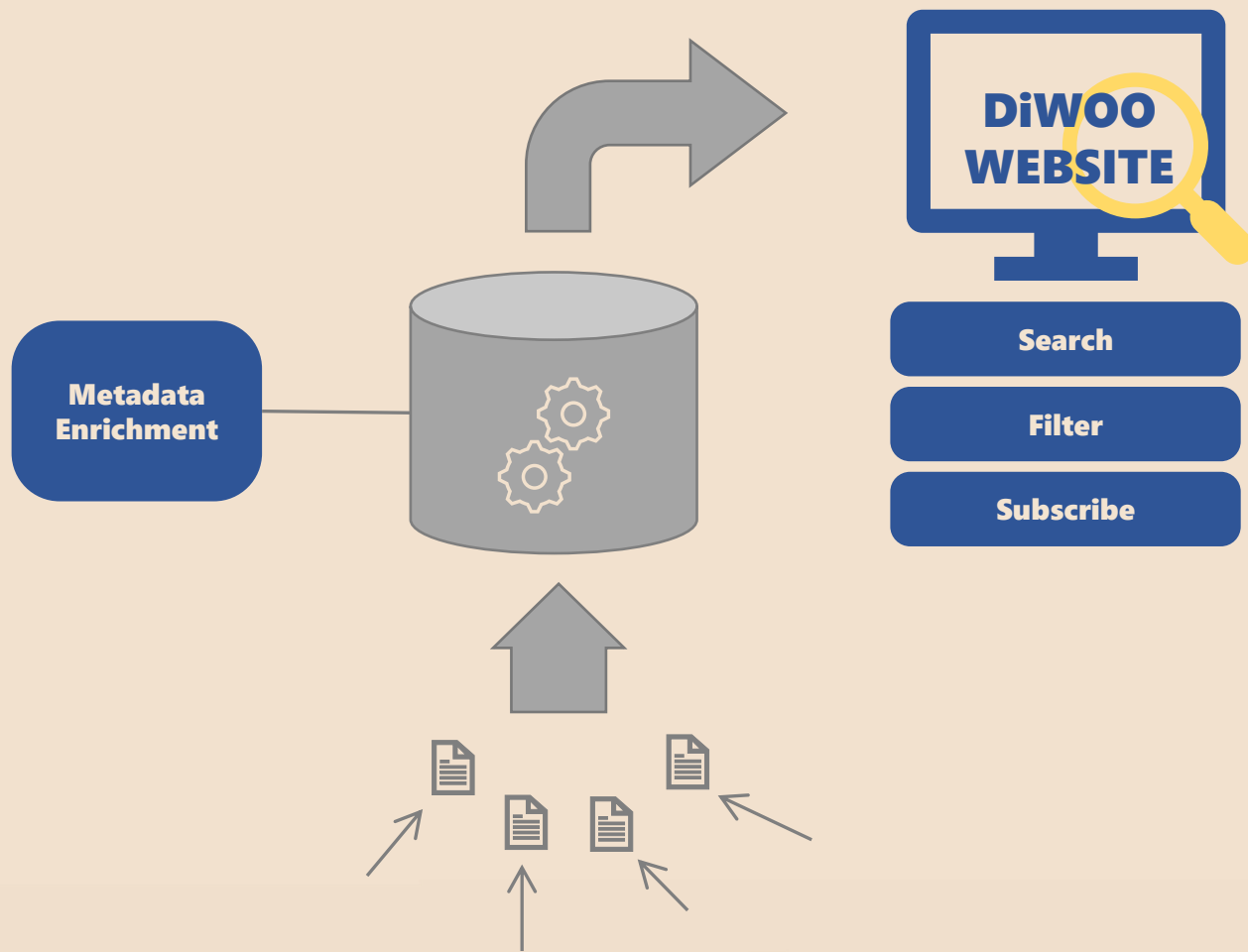
How many COVID cases per municipality?
How many citizens per municipality?



3

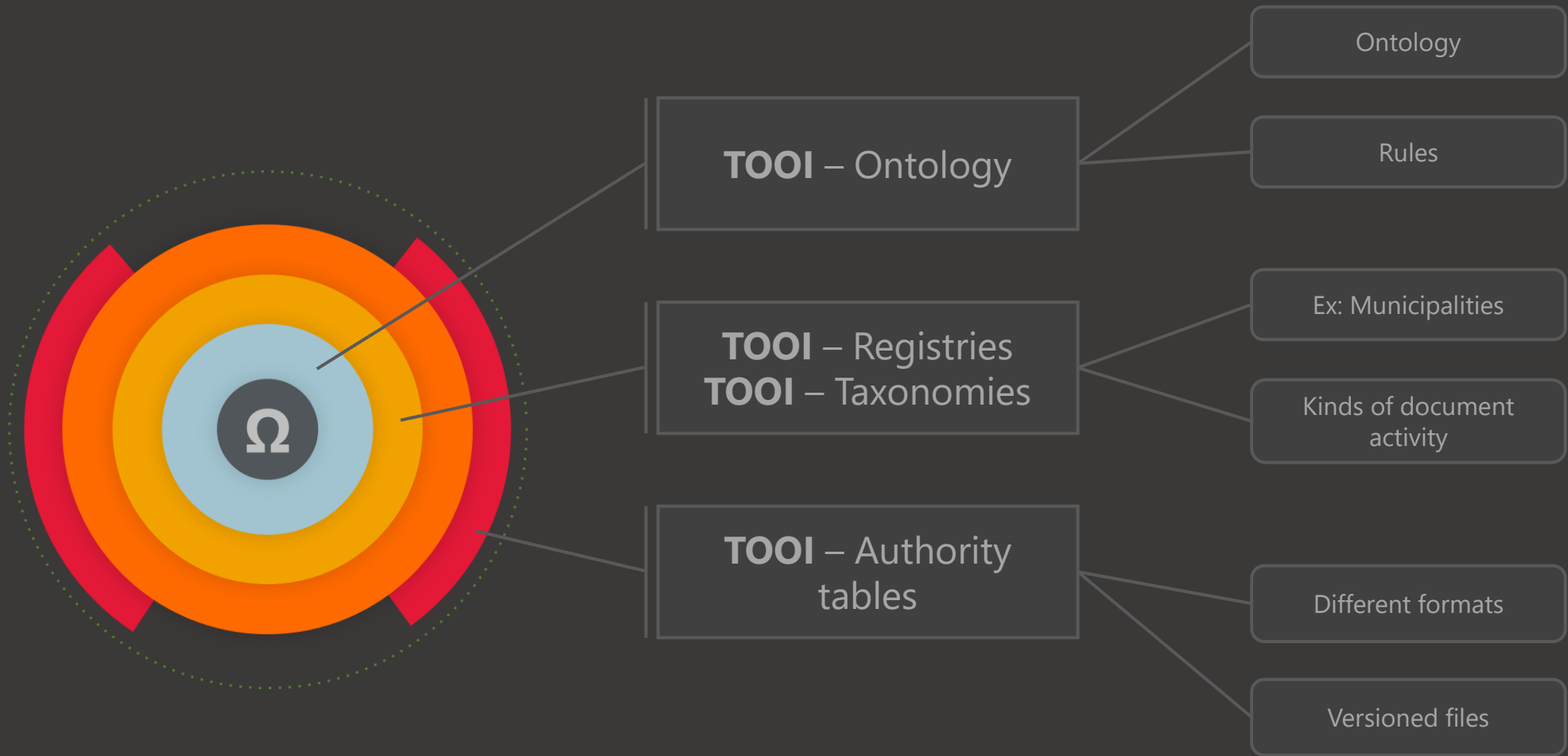
Chapter

1. The case of TOOI
2. Data federation with RDF



OPEN GOVERNMENT GOVERNMENT INFORMATION

- ❖ KOOP, publications office of the Netherlands (MinBZK)
- ❖ hings, not strings
Den Haag
's Gravenhage
- ❖ FAIR Principles
- ❖ Millions of documents, thousands of organizations



(Meta)data for Official Government Information: the TOOI Ontology and Knowledge Graph*

Jan Voskuil¹, Marc van Opijnen², Hans Overbeek², Theun Fleer² and Wessel Schollmeijer²

¹ Taxonic B.V., The Netherlands

² KOOP, Ministry of Interior and Kingdom Relations, the Netherlands

Abstract

The TOOI knowledge graph aims to achieve the FAIR objectives for official government information in the Netherlands. Its relevance extends beyond the immediate context in which it is conceived. This article presents the general characteristics of TOOI, how its constituting parts interrelate, and how its sustainability as a living standard is managed. It focuses on its core component, the TOOI ontology, and discusses some aspects of its design and development. It discusses how ODCM and OntoUML were applied, and reflects on practical aspects of the application of these methods.

Keywords

open government, (meta)data, ontoUML, ontology.

1. Introduction

TOOI [1] (acronym for 'Thesauri and Ontologies for Official Government Information') is a reference model in which authoritative information about public organizations and open government information is made available in a structured and machine-readable format for the purpose of coherence and findability of such information from various sources. Ultimately, TOOI's goal is to make such information FAIR [1]. This article focuses on the TOOI ontology in the context of the broader knowledge graph.

1.1. Problem statement

In today's complex and highly digitalized society, public transparency is of the utmost importance. Not only in complex crises, like the Covid-19 outbreak, but also in day-to-day life, lawyers, journalists, businesses, special interest groups and the general public at large, but also public organizations themselves, increasingly need coherent official documents and public data from a variety of sources. For instance, all those stakeholders have to be

FOIS 2024, Ontology showcase, July 15–19, 2024, Enschede, The Netherlands

* Part of the material in this paper is also available in Dutch as part of the TOOI documentation, see references.

• Corresponding author.

✉ jan.voskuil@taxonic.com (J. Voskuil); {marc.opijnen, hans.overbeek, theun.fleer, wessel.schollmeijer}

@koop.overheid.nl

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

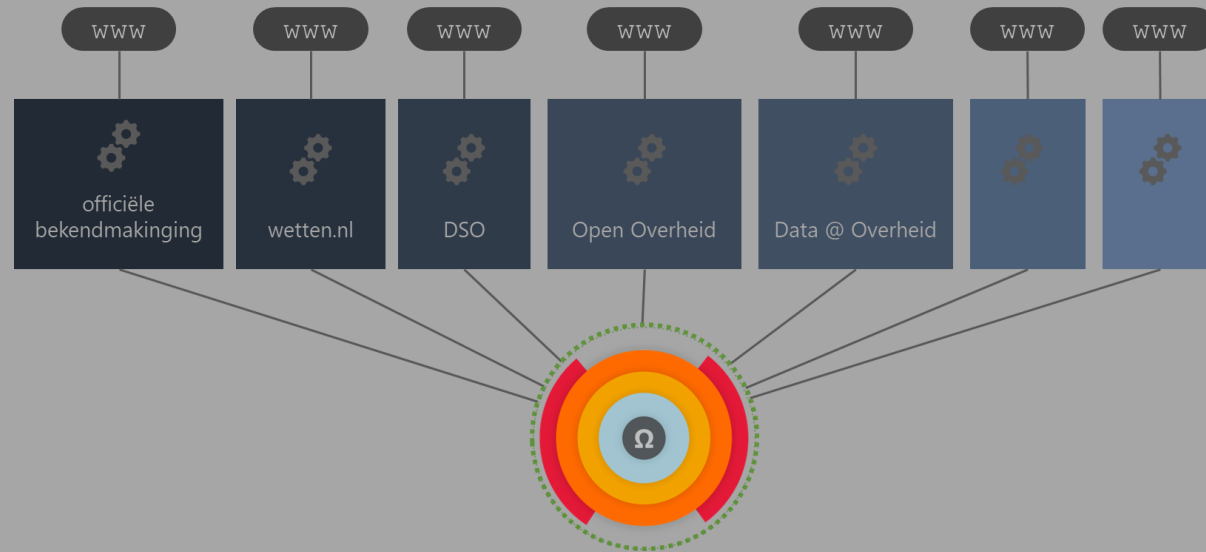
Te presenteren op FOIS 2024

standaarden.overheid.nl/tooi

The screenshot shows the 'Overheid.nl' website interface. At the top, there is a navigation bar with 'Overheid.nl', 'Home', 'Waardelijsten', and 'Documentatie'. Below the navigation bar, the page title is 'Overzicht van TOOI Waardelijsten'. A sub-header reads 'Selecteer in de onderstaande lijst een waardelijst om de inhoud te raadplegen of te downloaden.' Below this, there are two columns of links. The left column includes links for 'Betrokkenheid', 'Filetypes', 'KOOP-systemen', 'PLOOI documenthandelingen', 'PLOOI documentsoorten aanlevering', 'PLOOI filetypes aansluitvoorwaarden', 'Publicatiesoorten', 'Register Caribische openbare lichamen compleet', 'Register gemeenten compleet', 'Register ministeries compleet', 'Register overige overheidsorganisaties compleet', 'Register provincies compleet', 'Register waterschappen compleet', 'STOP bestuursorganen', 'STOP proceduresoorten', 'STOP stappen besluitvorming definitief besluit', 'STOP worktypes', 'Thema-indeling voor Officiële Publicaties (TOP-lijst)', 'Verdragstemas', 'Wep activiteiten invoer centraal', 'Wep plonsoorten invoer', 'Wep rubrieken invoer', 'Wep rubrieken invoer centraal', and 'Wep rubrieken volledig'. The right column includes links for 'Documentrelatie', 'Filetypes DCAT AP DQNL', 'Overheidsinformatie', 'PLOOI documentsoorten', 'PLOOI documentsoorten portaal', 'Publicatiebladen', 'Rechtsgebieden Basis wettenbestand', 'Register Caribische openbare lichamen op peildatum', 'Register gemeenten op peildatum', 'Register ministeries op peildatum', 'Register overige overheidsorganisaties op peildatum', 'Register provincies op peildatum', 'Register waterschappen op peildatum', 'STOP formaten van informatieobjecten', 'STOP regelingssoorten TPOD', 'STOP stappen besluitvorming ontwerpbesluit', 'Talen', 'Themas Basis wettenbestand', 'Wep activiteiten invoer centraal', 'Wep activiteiten volledig', 'Wep plonsoorten volledig', 'Wep rubrieken invoer beschikkingen zonder omgevingsvergunning', and 'Wep rubrieken invoer decentraal'.

USING TOOI

A DATA GOVERNANCE CHALLENGE



- > Non-invasive data governance
- > Slow uptake, little steps
- > Things, not strings, one field at a time


Long term goal:

Full use of the knowledge graph, in all its aspects



FAIR Data in Health Care

The case of KIK-V

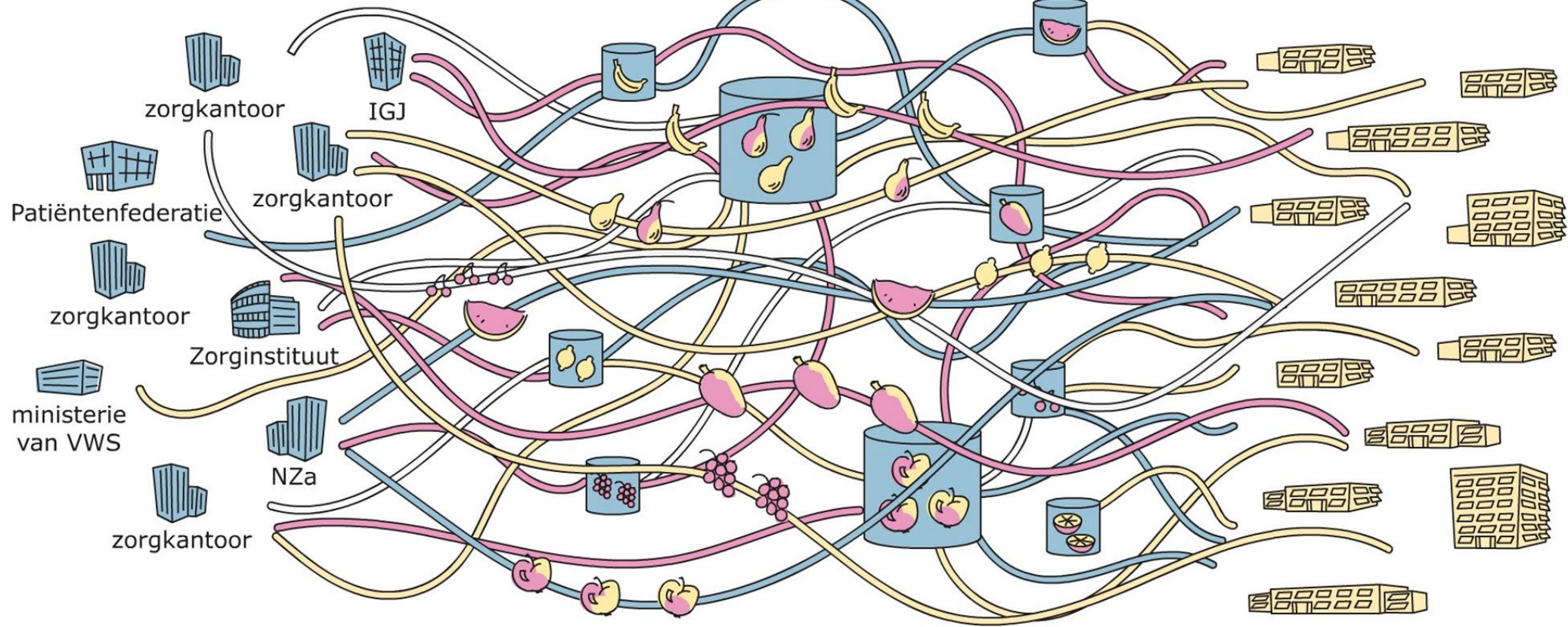


Ik wil dezelfde data
als die meneer,
maar dan met
mosterd ipv
ketchup

Probleem:
Tientallen ketenpartners willen data
van > 600 zorgaanbieders, ad-hoc

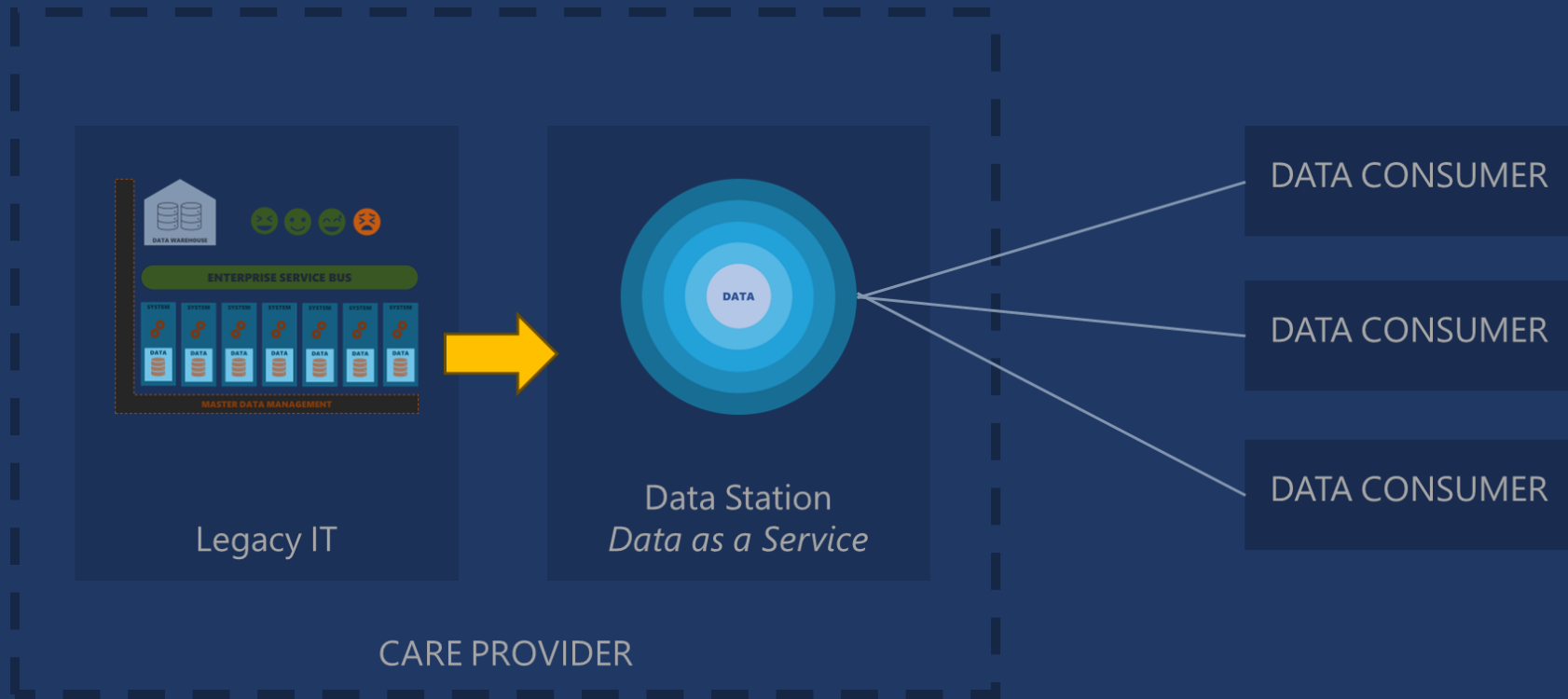
Chain parties

(Care) providers



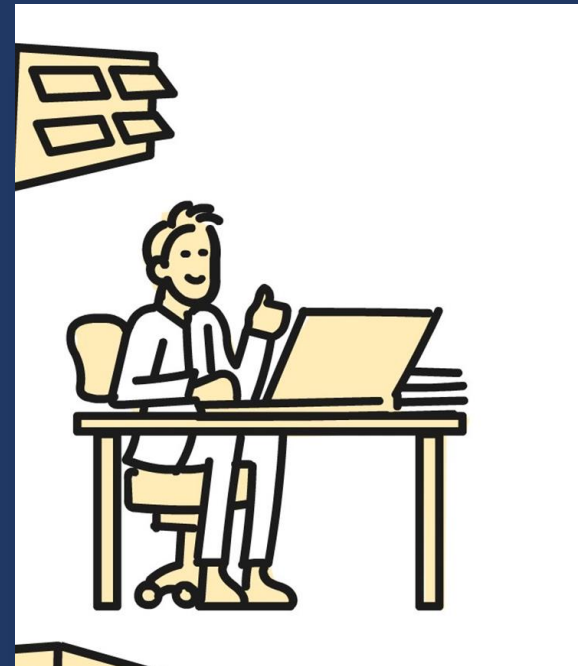
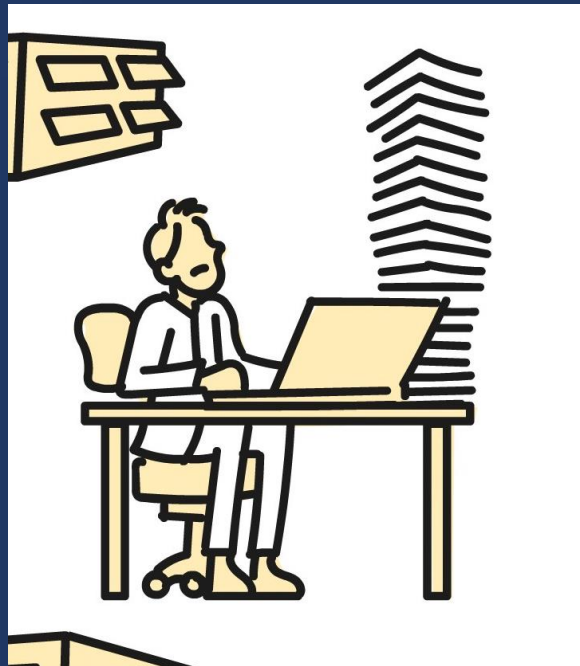
SOLUTION

DATA AS A SERVICE



RESULT

MORE RESOURCES FOR CARE



DATA GOVERNANCE

MEETING THE CHALLENGES

ONTOLOGY

Sufficiently complex, as simple as possible

STANDARD QUERIES

Governance body defining queries

STANDARD INDICATORS

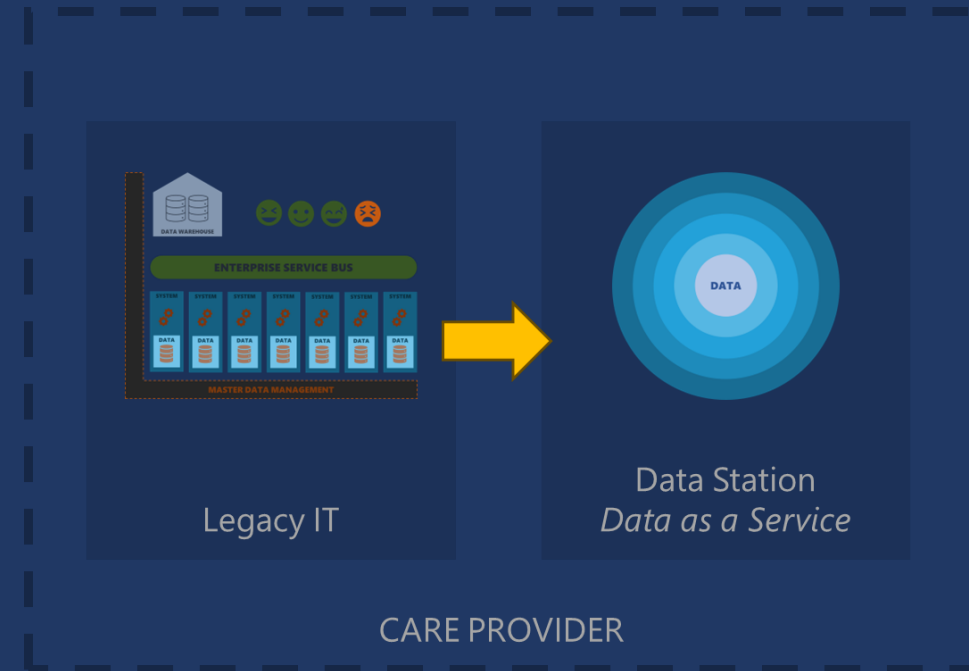
Governance body defining indicators

SECURITY & PRIVACY

Verifiable credentials, vetting

DATA TRANSFORMATION

Care providers are responsible, IT-providers help



4

Chapter

Data centrality and the
world of GIS



DATA GOVERNANCE

THE PROBLEM WITH GIS

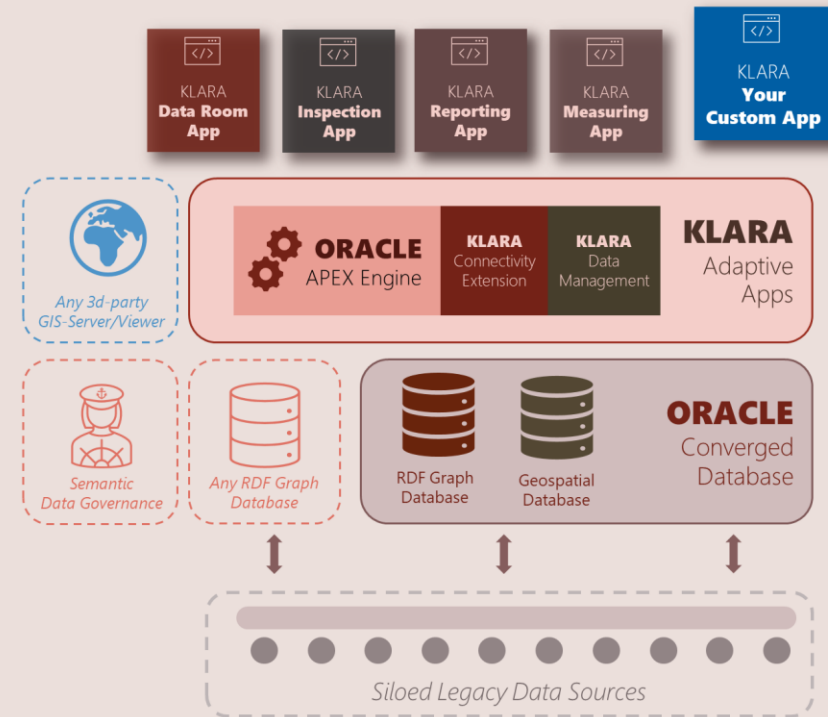
- > Data duplication from source to GIS
- > Adding datasets is expensive
- > Disparate data difficult to integrate
- > Editing on the map causes inconsistencies

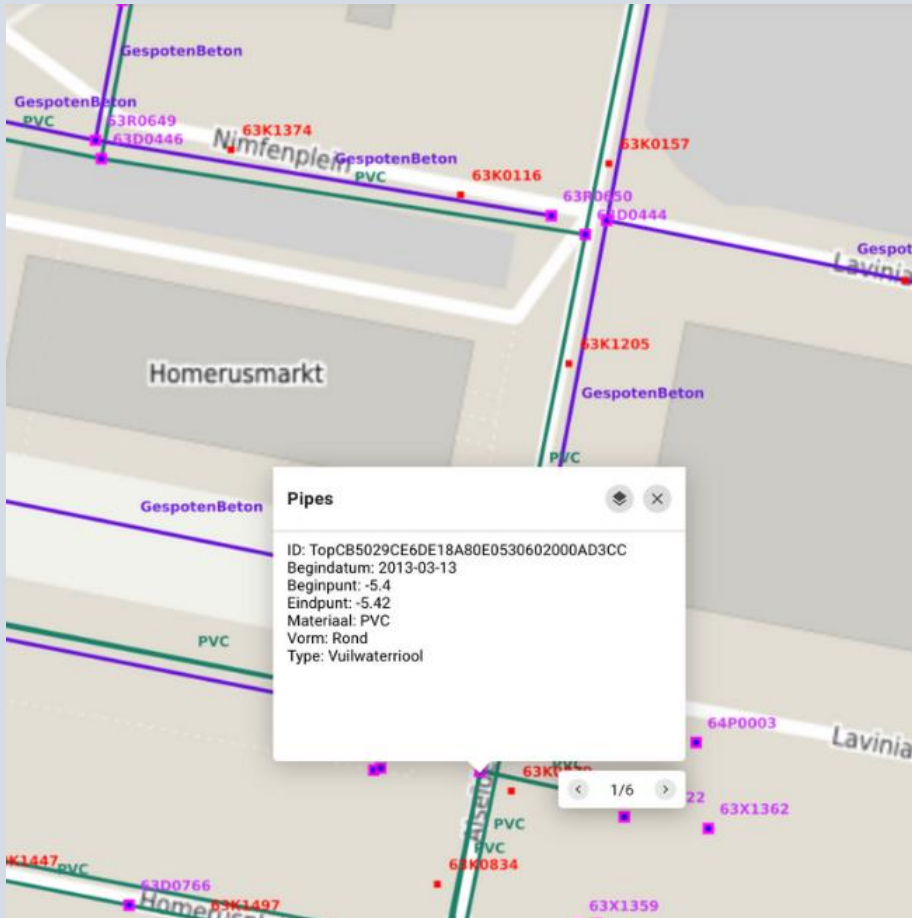


PROPOSITION

BASED ON GWSW-DATASETS...

- > A cloudbased managed service
- > Integrated management of registrative & geospatial data
- > Focussed on the domain of **asset management**.
- > Supports **low-code**
- > Builds on three core elements of the Oracle stack:
 - ❖ Oracle RDF Graph Database
 - ❖ Oracle Geospatial Database
 - ❖ Oracle Apex low-code application engine





FUNCTIONALITY

CORE FEATURES

Object registration

Klara functions as the central object registration (IMBOR)

Integrated data management

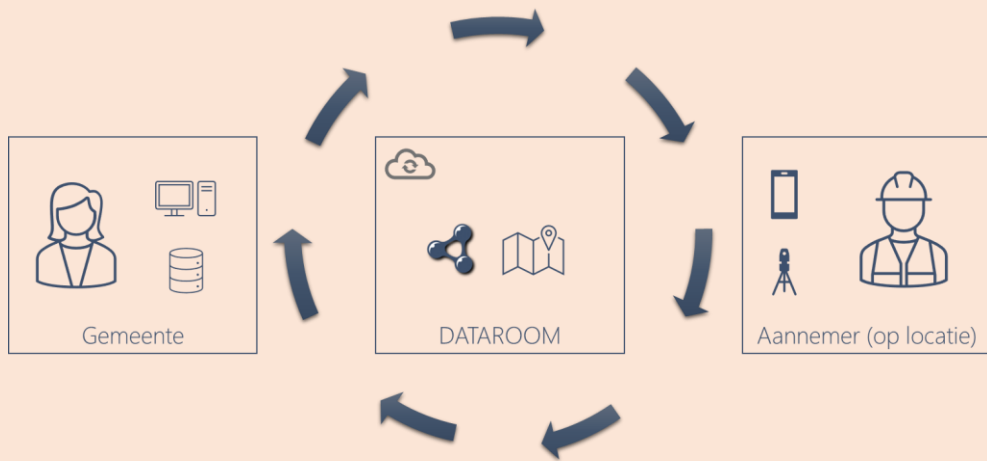
Through Klara, registrative passport data and geospatial data are managed as an integrated whole.

Object information available in a diversity of processes

In many processes, object data needs to be shared with zero loss of meaning

REVISIE ALS USE CASE

FUNCTIONALITEIT



DATA GOVERNANCE AND AI

CRUCIAL TAKE-AWAYS

1

AI IS NOT READY

Knowledge graphs help AI:

- > Better performance
- > Explainable AI

2

DATA IS NOT READY

Data centric architecture:

- > Consistent data across processes
- > Federated data

3

THE ORGANIZATION
IS NOT READY

Innovate data governance:

- > make data centricism a priority
- > learn RDF Graph technology



Thank you!

jan.voskuil@taxonic.com

manon.dijkstra@taxonic.com