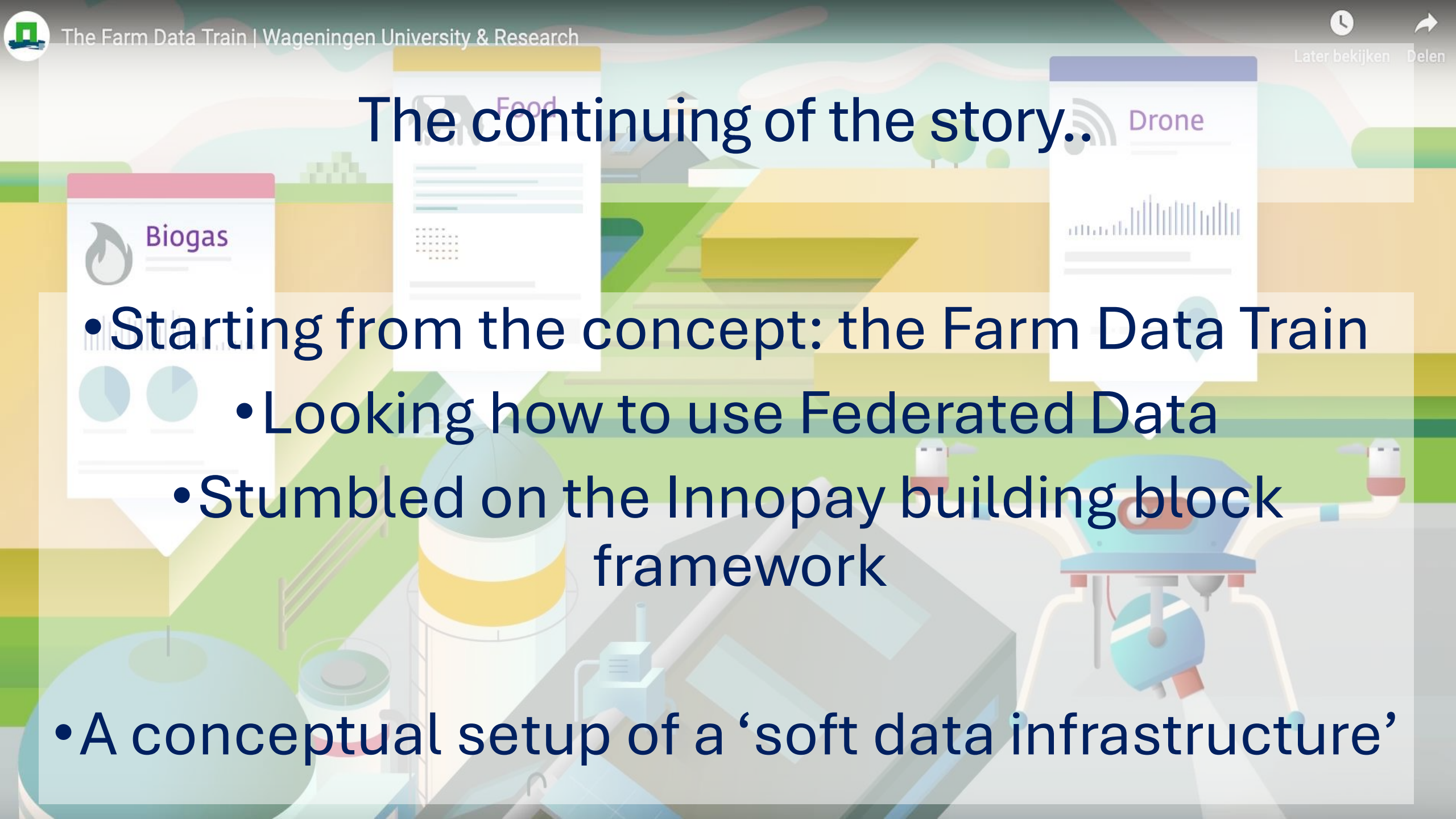

Towards a
common data infrastructure
Jandirk Bulens, Marc-Jeroen Bogaardt, Rob Lokers,
Claudia Kamphuis, Fedde Sijbrandij
©sdecoret/Adobe Stock

What did we do?

- Develop and improve smart, privacy conserving data infrastructure
  - Around topic of circular agriculture
  - Using farm generated data and farm and food safety sensors
  - Touching emerging issues around privacy conservation, data sovereignty, data protection
- Establishing connections between infrastructure components
  - Test bed for exploring connectivity, interoperability, reusability over WUR infrastructures and research domains
- Mapping to a generic infrastructural framework
- Assessment of WUR organisational perspectives on data architecture
- Compiling lessons learned and a roadmap for future evolution

How? → by connecting the Dots

- Farms Information Net (BIN)
  - Sample companies, data warehouse, business intelligence stack
- Farmmaps
  - Farm data space, management & advicemanagement information system
- AgroDataCube
  - Open data, legislative/satellite/sensor data service
- Farm Data Safe
  - Data sovereignty & privacy protection, personal data vault
- Methane Data Lake
  - On-farm sensor data streams, data lake, real-time
- Food Fraud Data (RASFF/EMI)
  - Sensitive/protected data, federated learning
- Fresh Water Fish
  - Sectoral interoperability, information model

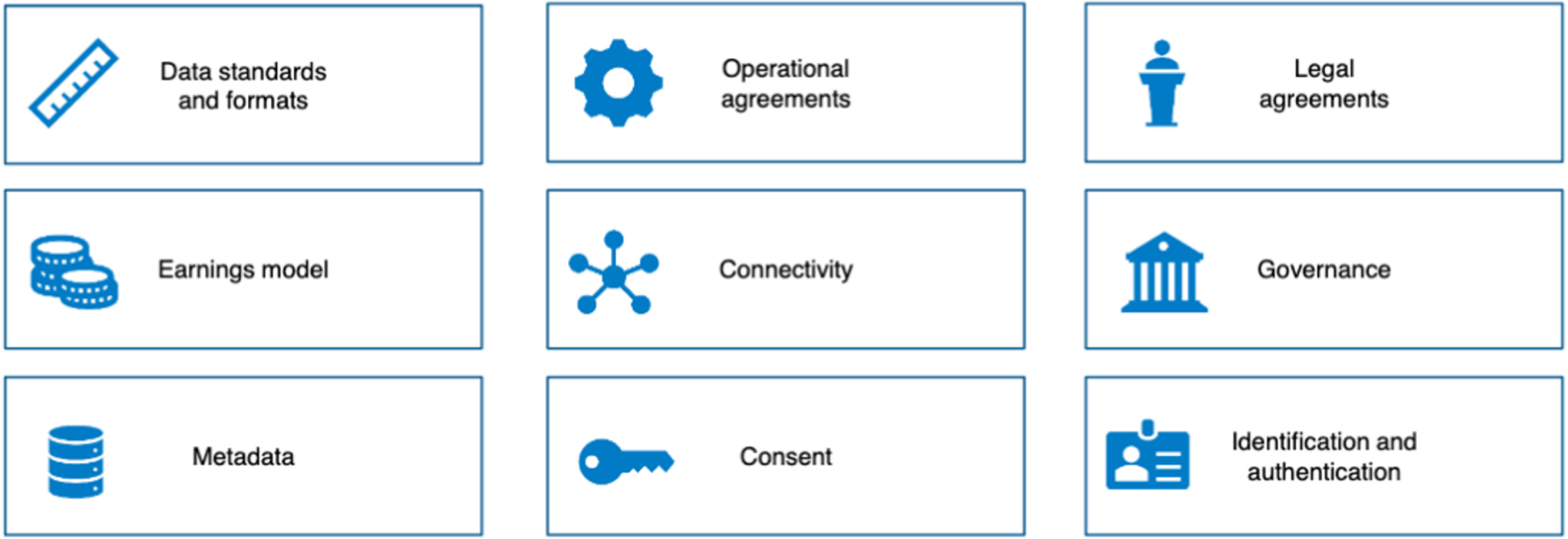


The continuing of the story..

- Starting from the concept: the Farm Data Train
  - Looking how to use Federated Data
  - Stumbled on the Innopay building block framework

- A conceptual setup of a ‘soft data infrastructure’

Infrastructural building blocks framework



Source: Survey and benchmark data sharing initiatives, INNOPAY analysis

What about you?

What standards do you use?

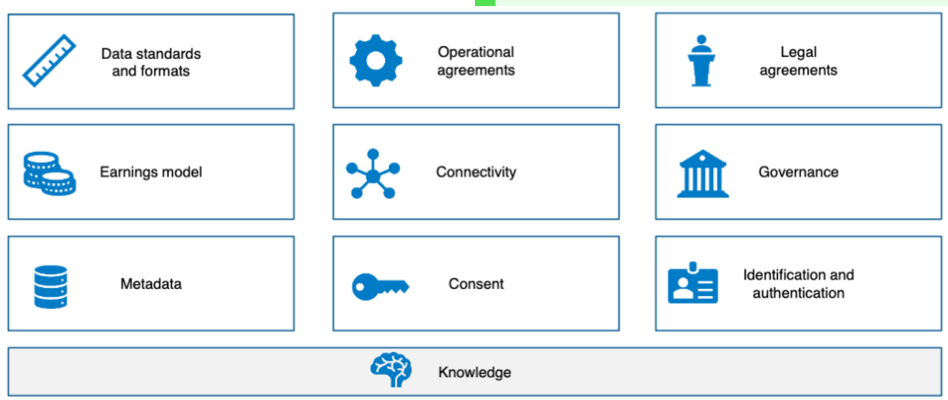
What license do you use?

Do you provide access?

Do you create Metadata?

For the data collection? - For the data structure? - Do you publish metadata? .. and how?

How do you authorise users?

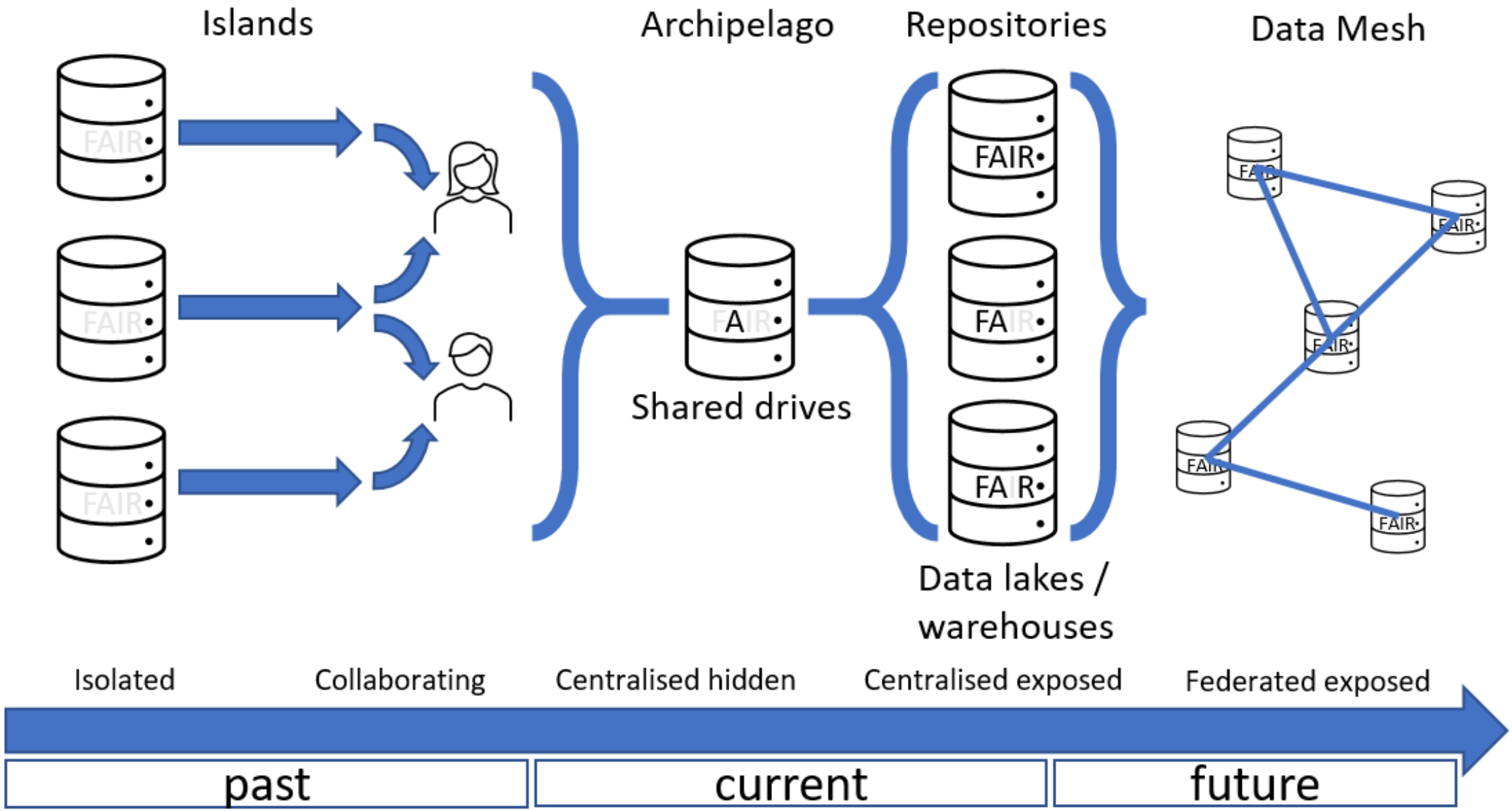


Source: Survey and benchmark data sharing initiatives, INNOPAY analysis

Building block	Initial /experimental phase	Data standards	Earnings model	Metadata	Operational agreements	Connectivity	Consent	Legal agreements	Governance	Identification & authentication
Data source										
Farm Information Net (BIN)										
Farmmaps										
AgroDataCube (ADC)										
Farm Data Safe	yes									
Methane Data Lake	yes									
Food Fraud Data (RASFF/EMI)										
Fresh Water Fish										

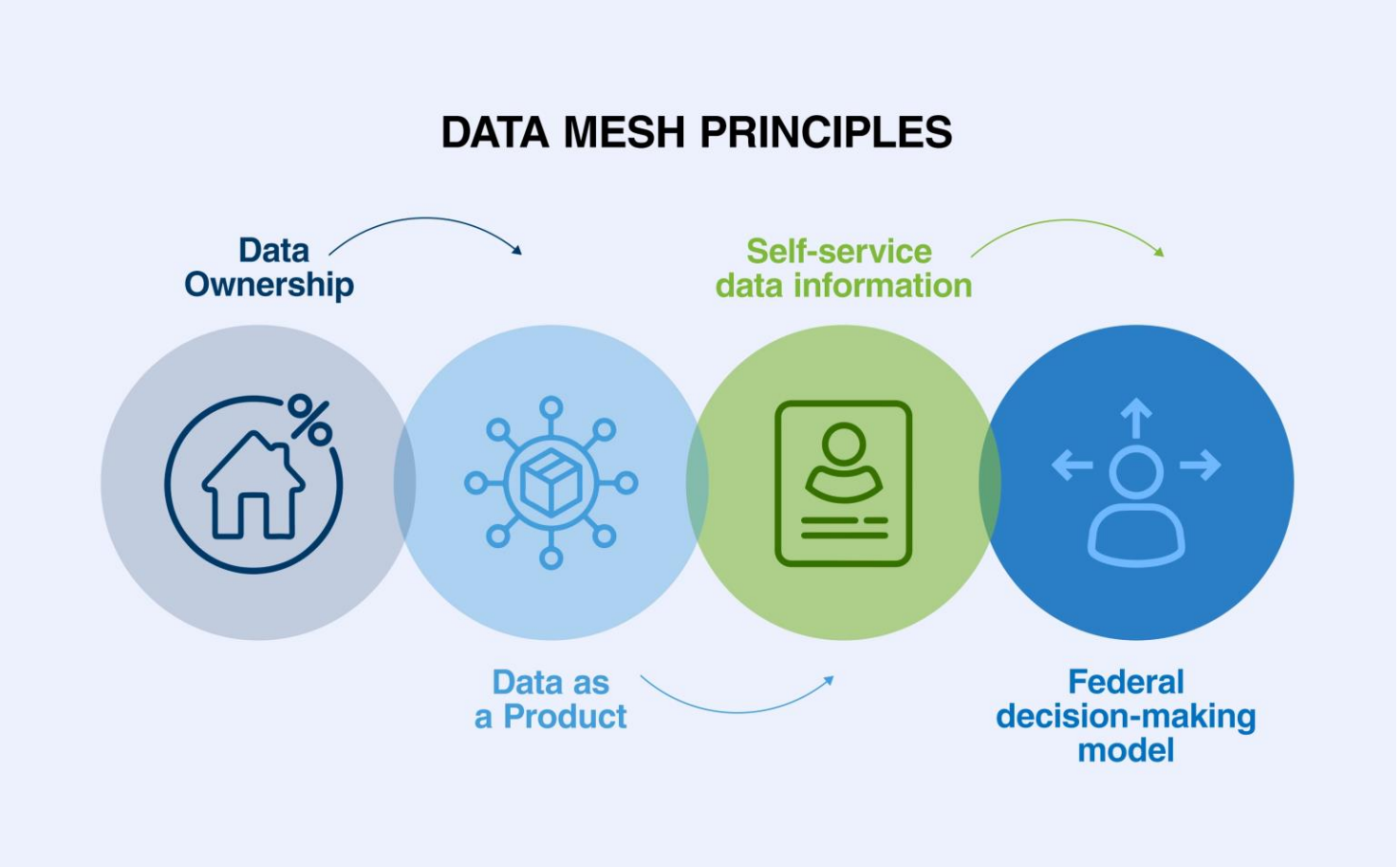
Findings

- What is still immature, or even lacking, are the standards that relate to semantics.
- Regarding the earnings model: also scientific data even from ongoing research can be very valuable for re-use.
- Metadata is often misunderstood. A general metadata standards is Dublin Core Metadata Initiative (DCMI). More specific for the WUR, PURE
- It is good to understand that the 9-block framework was in most cases not known when the data sources were compiled.
- within WUR there is a strong notion to make data FAIR, meaning a lot of building blocks already got attention.
- providing interoperability is the most difficult aspect of 'FAIRifying' data to compile proper metadata.
  - Happily enough we see that controlled vocabularies and ontologies are used more and more.


Islands
Archipelago
Repositories
Data Mesh
FAIR
FAIR
FAIR
A
Shared drives
FAIR
FA
FA R
Data lakes /
warehouses
Isolated
Collaborating
Centralised hidden
Centralised exposed
Federated exposed
past
current
future

Data mesh principles

- Data Product versus
- Data as a Product



Data mesh pioneer Zhamak Dehghani talks about *calling for a new principle, self-serve data infrastructure as a platform to enable domain autonomy.*

All data elements

(basically, make data FAIR)

Recommendations [General]

- Disconnect data form applications, in contrast to data owners now tempted to build applications on top of the data and then develop a complete infrastructure (‘data as a product’ concept)
- Design a data exchange data infrastructure based on predetermined requirements for WUR
- Raise awareness on sharing data by (better) exposing a common vision providing support on all aspects as also shown in the data sharing Infrastructure framework
- Strengthen and improve the position of Data Stewards to raise awareness and provide better support on a ‘sharing data infrastructure’

Recommendations [all Building Blocks]

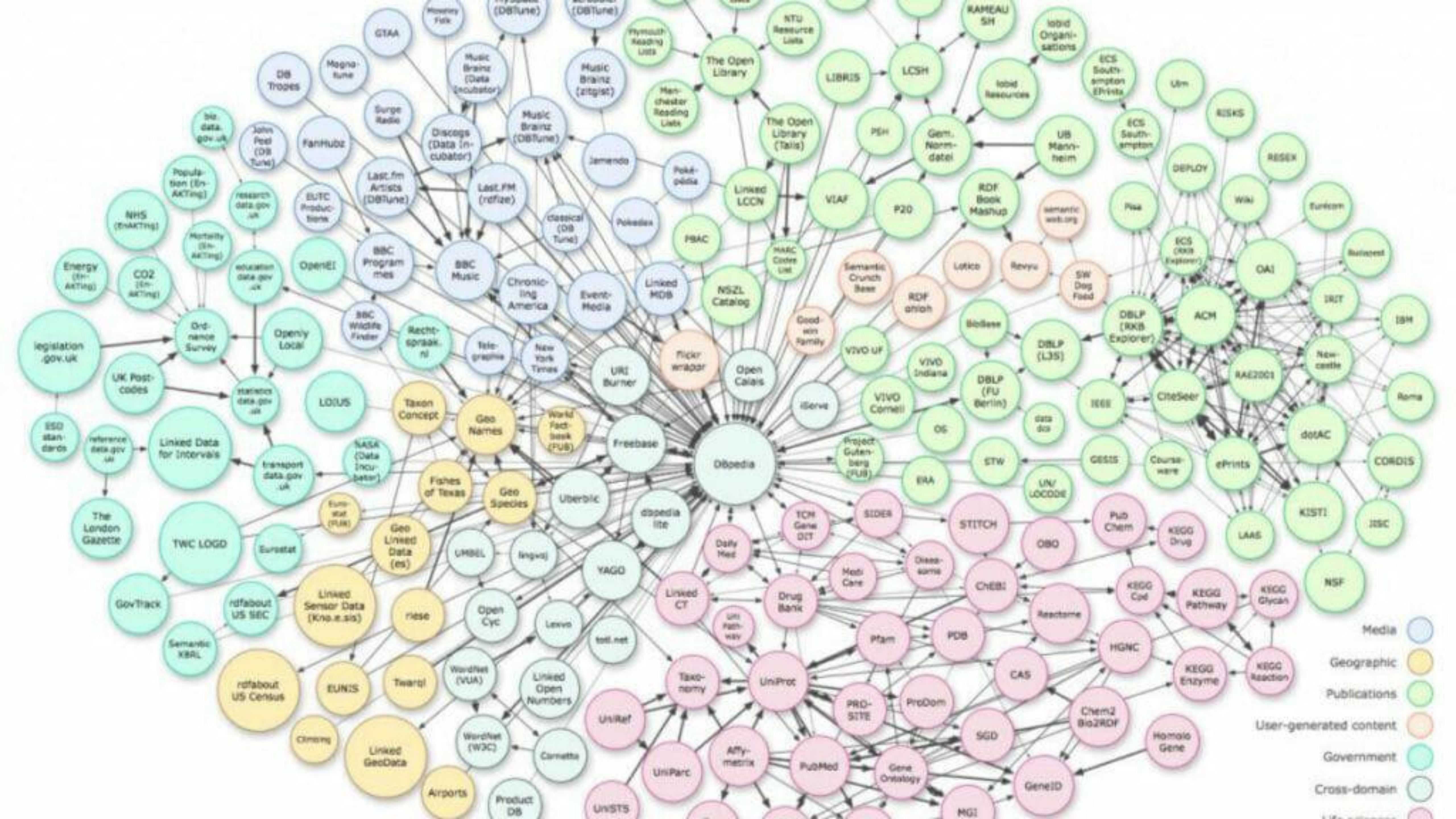
- Closely follow developments in shared data infrastructures taking place internationally
- At some point we need to decide to stop local initiatives and move to European level and phase out our legacy applications / infrastructures

Recommendations [Overall/Governance]

- We recommend to promote and support the process by more consciously following a framework for a soft infrastructure, such as the nine building blocks or the framework for the EU-dataspaces
- When implementing provide a unified (WUR) infrastructure, ready to use
- Our overarching recommendation would be to define governance more explicit related to the data infrastructure (a role of Chief Data Officer (CDO@WUR) possibly assigned to one of the general directors?)

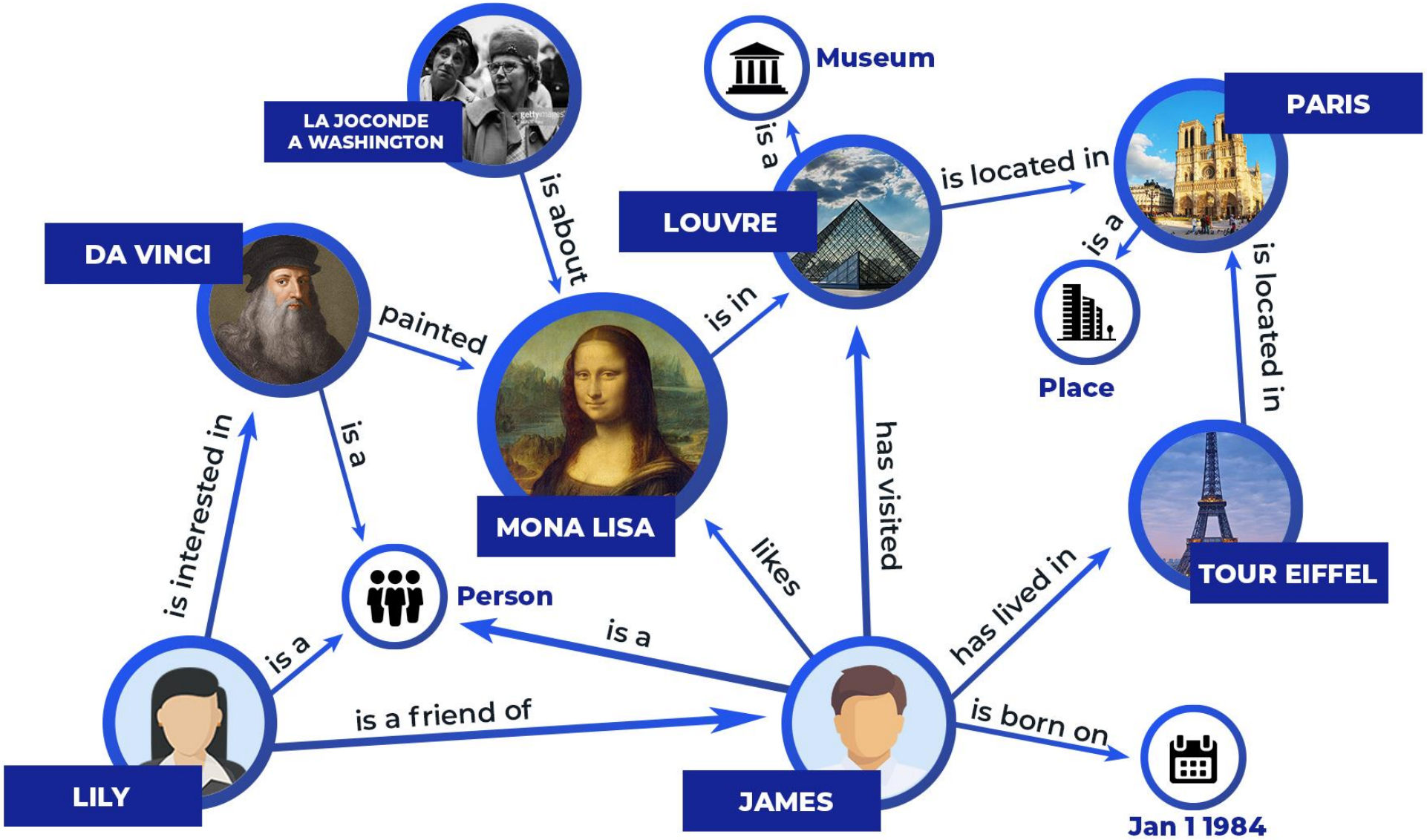
Recommendations [BB Metadata]

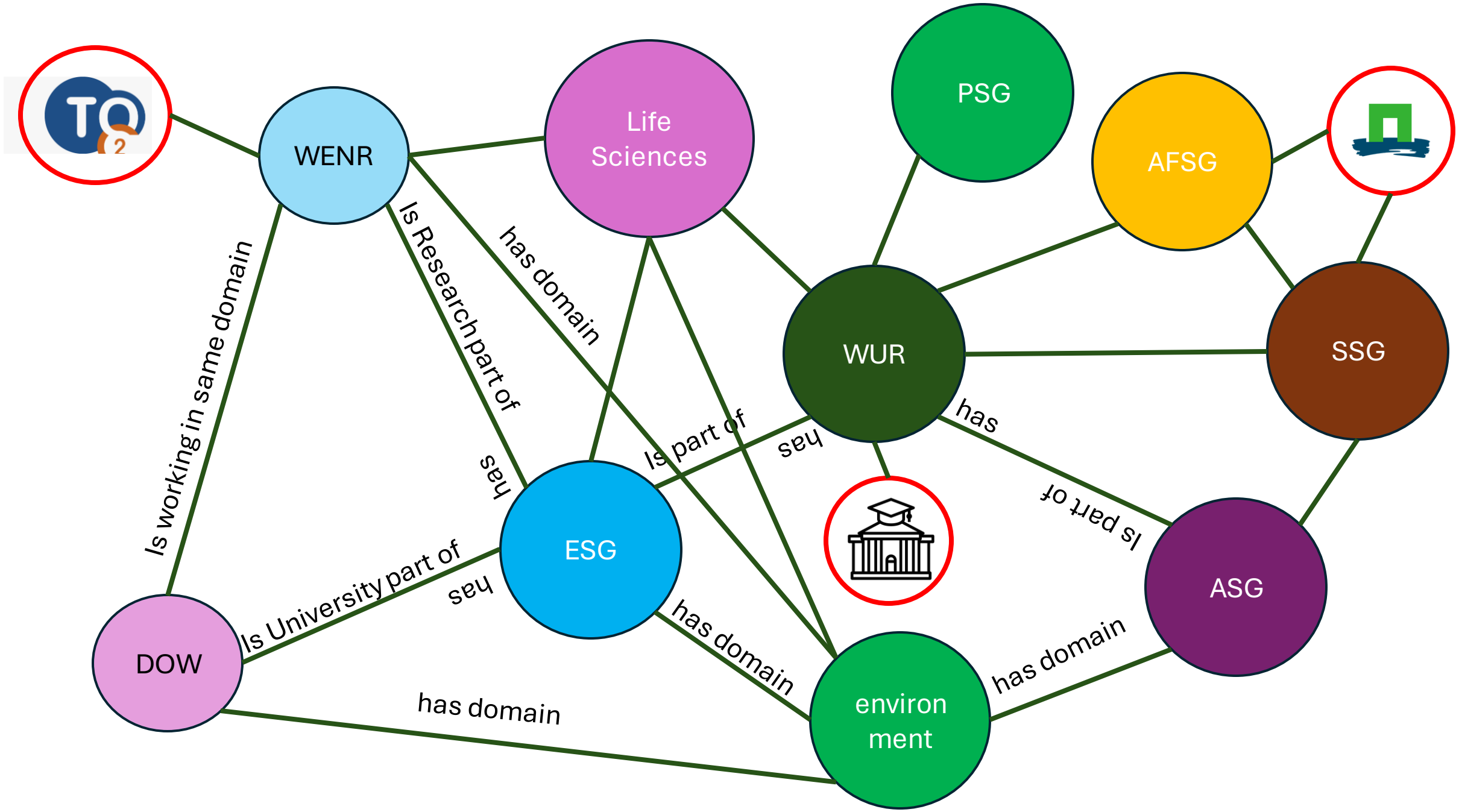
- We recommend to make use of semantic standards and create proper schemas with clear definitions and relationships between objects (ontologies) by using as much as possible existing commonly accepted ontologies in combination with controlled vocabularies and/or thesauri.
- We recommend to use a common WUR standard to create Metadata that combines the best of the current standards in use and fit the requirements of the WUR.
- Promote the idea of developing a WUR Knowledge graph. Start a programme on this at corporate level (KB?)



GTAA
DB Tropes
Magna-tune
Music Brainz (Data Incubator)
Music Brainz (zitgist)
Surge Radio
FanHubz
John Peel (DB Tune)
Discogs (Data Incubator)
Music Brainz (DBTune)
Jamendo
Last.fm Artists (DBTune)
Last.FM (rdfize)
Poké-pédia
EUTC Productions
classical (DB Tune)
Pokedex
BBC Programmes
BBC Music
Chronicling America
Event-Media
Linked MDB
BBC Wildlife Finder
Recht-spraak.nl
Tele-graphis
New York Times
URI Burner
flickr wrappr
Open Calais
bis.data.gov.uk
Population (EnAKTing)
research data.gov.uk
NHS (EnAKTing)
Mortality (EnAKTing)
Energy (EnAKTing)
CO2 (EnAKTing)
education data.gov.uk
OpenEI
legislation.gov.uk
Ordnance Survey
Openly Local
UK Post-codes
statistics data.gov.uk
LOIUS
ESD standards
reference data.gov.uk
Linked Data for Intervals
transport data.gov.uk
NASA (Data Incubator)
The London Gazette
TWC LOGD
Eurostat
GovTrack
rdfabout US SEC
Semantic XBRL
rdfabout US Census
Taxon Concept
Geo Names
World Factbook (FUB)
Fishes of Texas
Geo Species
Eurostat (FUB)
Geo Linked Data (es)
Linked Sensor Data (Kno.e.sis)
riese
EUNIS
Twarql
Climbing
Linked GeoData
Airports
Freebase
DBpedia
Uberblic
dbpedia lite
UMBEL
lingvoj
YAGO
Open Cyc
Lexvo
totl.net
WordNet (VUA)
Linked Open Numbers
WordNet (W3C)
Cornetto
Product DB
Plymouth Reading Lists
NTU Resource Lists
The Open Library
RAMEAU SH
lobid Organisations
Manchester Reading Lists
LIBRIS
LCSH
lobid Resources
ECS Southampton EPrints
Ulm
The Open Library (Talis)
PSH
Gem. Normdatei
UB Mannheim
ECS Southampton
RISKS
RESEX
Linked LCCN
VIAF
P20
RDF Book Mashup
DEPLOY
Wiki
Pisa
Eurécom
PBAC
MARC Codes List
semantic web.org
ECS (RKB Explorer)
Budapest
Semantic Crunch Base
Lotico
Revyu
SW Dog Food
OAI
NSZL Catalog
RDF ohloh
DBLP (RKB Explorer)
ACM
IRIT
IBM
Good-win Family
BioBase
DBLP (L3S)
VIVO UF
VIVO Indiana
New-castle
RAE2001
DBLP (FU Berlin)
iServe
VIVO Cornell
IEEE
CiteSeer
Roma
data dcs
OS
dotAC
Project Gutenberg (FUB)
ERA
STW
GESIS
Course-ware
ePrints
CORDIS
UN/LOCODE
KISTI
JISC
LAAS
NSF
TCM Gene DIT
SIDER
STITCH
Pub Chem
KEGG Drug
Daily Med
OBO
Medi Care
Disea-some
ChEBI
KEGG Cpd
KEGG Pathway
KEGG Glycan
Linked CT
Drug Bank
Reactome
Uni Path-way
Pfam
PDB
HGNC
KEGG Enzyme
KEGG Reaction
CAS
Taxo-nomy
UniProt
PRO-SITE
ProDom
Chem2 Bio2RDF
UniRef
SGD
Homolo Gene
UniParc
Affy-metrix
PubMed
Gene Ontology
GeneID
UniSTS
MGI
Media
Geographic
Publications
User-generated content
Government
Cross-domain

A knowledge graph?




TO2
WENR
Life Sciences
PSG
AFSG
SSG
WUR
ESG
ASG
DOW
environ ment
Is working in same domain
Is Research part of
has domain
has
Is part of
has
Is part of
Is University part of
has
has domain
has domain
has domain

What is next?

- The report is ready to be published
- Identify key elements of a soft infrastructure for WUR
- Interview the WCDS projects and key researchers on the key elements
  - Current experiences?
  - How to use for future work/research?
  - Prioritise the key elements
- Detail the roadmap for WUR
  - Related to the digital strategy
  - Follow up in KB and WUR programmes


©sdecoret/Adobe Stock