



# FAIR@Scale - lessons learnt

Gareth Murphy

Data Integration and Ontology, Novo Nordisk

9/2/2021

# Novo Nordisk® at a glance

Novo Nordisk is a leading global healthcare company, founded in 1923 and headquartered in Denmark.

Our purpose is to drive change to defeat diabetes and other serious chronic diseases such as obesity and rare blood and endocrine disorders.

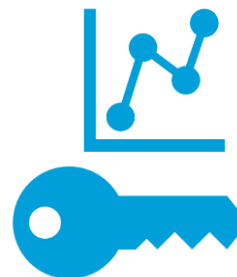
We do so by pioneering scientific breakthroughs, expanding access to our medicines and working to prevent and ultimately cure disease.



# What is FAIR data?



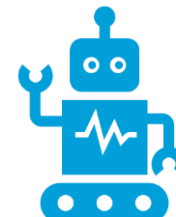
Findable – you can find it on the intranet



Accessible – you can log in and read/plot the data



Interoperable – can read with non-custom software



Reusable – metadata describe it “well enough” so others & AI can use

# Why FAIR?



In order to be ready for machine learning/AI applications, such as AlphaFold, we need a large amount of harmonized training data

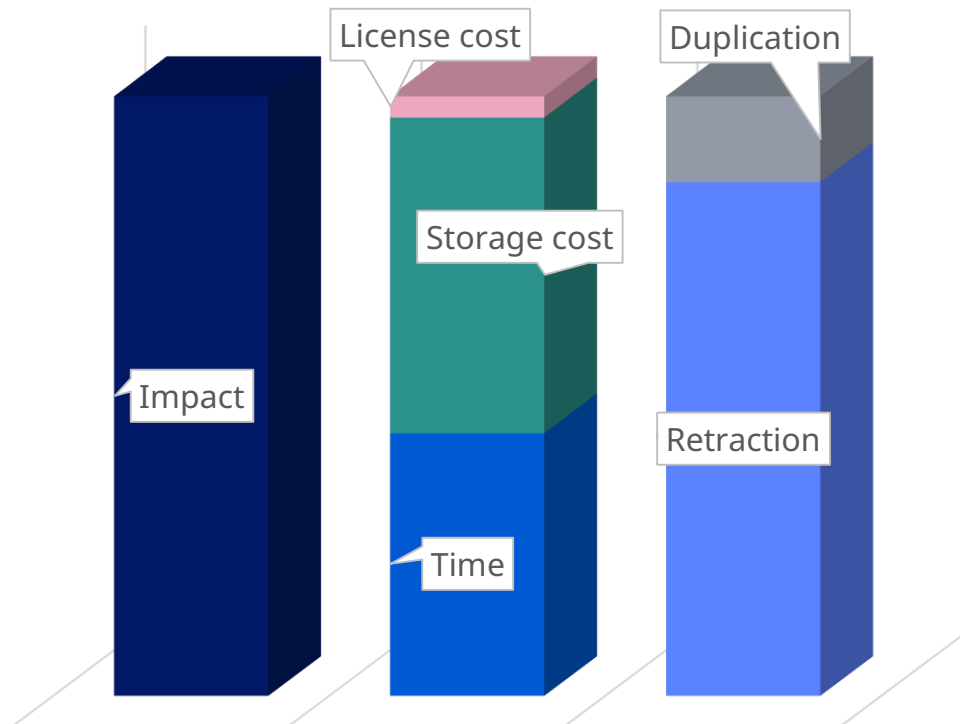


We need to capture negative results as well as positive results



Mitigate risk of redoing experiments

# Cost of FAIR@Source



## Cost of not being FAIR

\*European Commission

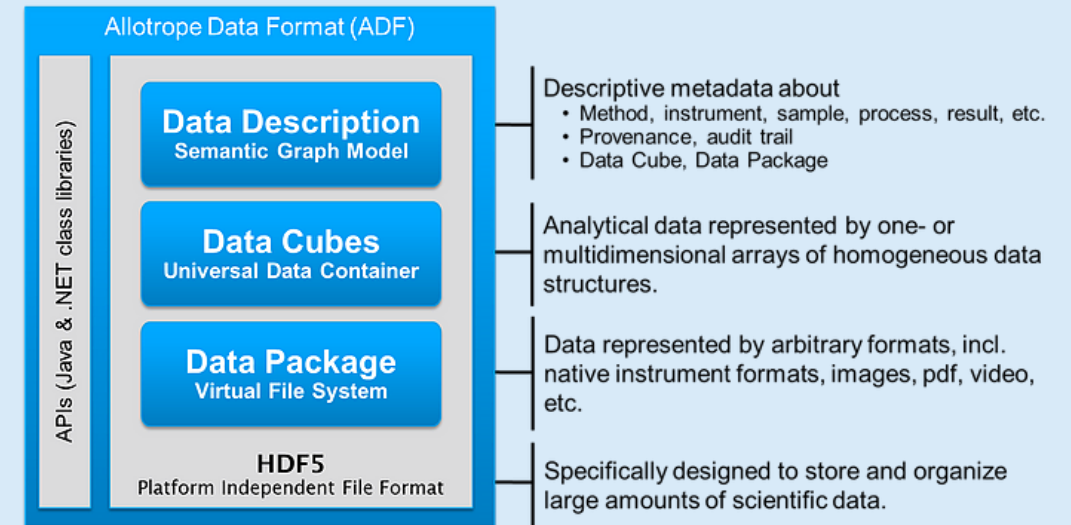
## Investment required

- Infrastructure
  - FAIR compliant applications
  - Services and Interfaces
- Culture
  - Governance team
  - Change management
- Policies
  - Standards & protocols
  - Persistent Identifiers
- Legacy data Harmonisation

FAIR ecosystem =  
culture + infrastructure

# FAIR culture –change management

- No more “my lab, my data”
- Data is a shared asset which needs to be saved, preserved and reused, harmonized, integrated
- Example - Interoperability:
- Currently all instruments use different data formats - 150+ current bio data formats –
- Can we convert to one self-describing multi-instrument format with rich metadata Allotrope Foundation



# FAIR assessment



- Interview of teams across R&ED for FAIR assessment.
- Find out how people can find their data, access – overall experience



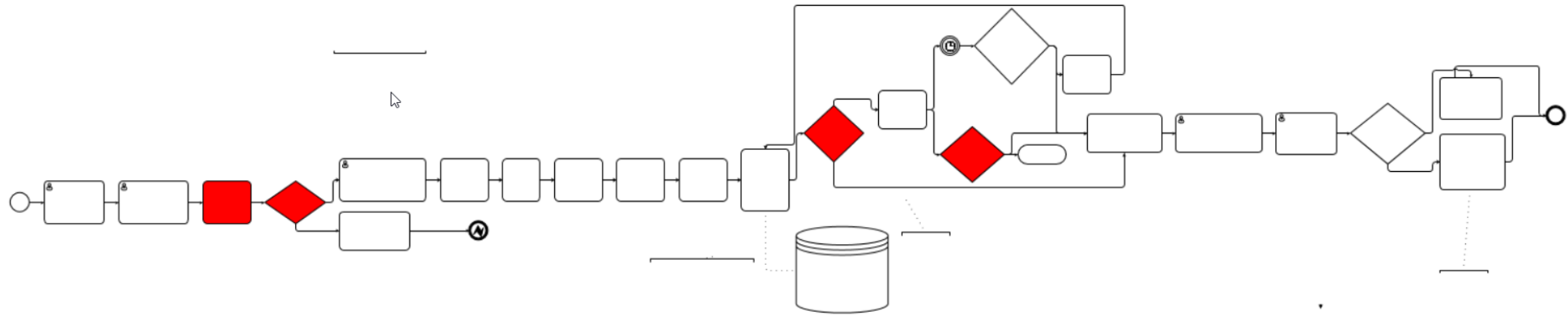
# FAIR Assessment

- How FAIR is our data?
- How widespread are use of PIDs, ontologies, controlled vocabularies?

Dept	Findability	Accessibility	Interoperability	Reusability
1	Green	Green	Yellow	Green
2	Yellow	Yellow	Yellow	Yellow
3	Green	Yellow	Green	Yellow
4	Green	Yellow	Orange	Green
5	Green	Red	Yellow	Green
6	Green	Green	Yellow	Yellow
7	Yellow	Green	Yellow	Red
8	Green	Orange	Red	Red
9	Orange	Green	Red	Green
10	Green	Red	Green	Green
11	Red	Green	Red	Yellow
...	Red	Orange	Yellow	Green

# Curathons

- What metadata is used and where, how can it be preserved across the data pipeline



# Heterogeneous datasets – harmonizing with NLP

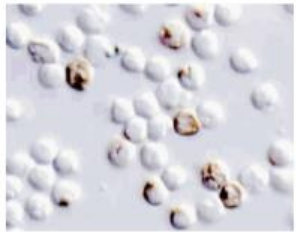
# FAIR @source, @scale – stakeholder engagement

Product name	Anti-ACAT1 antibody [EPR10359]
Description	Rabbit monoclonal [EPR10359] to ACAT1
Host species	Rabbit
Tested applications	Suitable for: WB, IP, IHC-P Unsuitable for: Flow Cyt or ICC/IF
Species reactivity	Reacts with: Human Predicted to work with: Mouse, Rat ▲
Immunogen	Synthetic peptide corresponding to Human ACAT1 aa 350-450. Database link: <a href="#">P24752</a>
Positive control	Human fetal liver, SW480, Jurkat, HepG2, (C) SKBR-3, THP-1 and Human fetal kidney lysates, Human heart & liver tissue, immunoprecipitation pellet from Human fetal liver lysate
General notes	<p>This product is a recombinant monoclonal antibody, which offers several advantages including:</p> <ul style="list-style-type: none"> <li>- High batch-to-batch consistency and reproducibility</li> <li>- Improved sensitivity and specificity</li> <li>- Long-term security of supply</li> <li>- Animal-free production</li> </ul> <p>For more information <a href="#">see here</a>.</p> <p>Our RabMAb® technology is a patented hybridoma-based technology for making rabbit monoclonal antibodies. For details on our patents, please refer to <a href="#">RabMAb® patents</a>.</p> <p>The Life Science industry has been in the grips of a reproducibility crisis for a number of years. Abcam is leading the way in addressing the problem with our range of recombinant monoclonal antibodies and knockout edited cell lines for gold-standard validation.</p> <p>One factor contributing to the crisis is the use of antibodies that are not suitable. This can lead to misleading results and the use of incorrect data informing project assumptions and direction. To help address this challenge, we have introduced an application and species grid on our primary antibody datasheets to make it easy to simplify identification of the right antibody for your needs.</p>

x8 pages



# FAIR @source, @scale – stakeholder engagement

Product name	Product name	Host	Product Information
Description	Description	Clonality	Material Number: 559064
Host species	Host species	Isotype	Alternate Name: Tnf, Tnfa; TNF- $\alpha$ ; Tnfifa; Tnfa2; TNFSF2; Cachectin; DIF
Tested application	Tested application	Application	Size: 0.25 mg
Species reactivity	Species reactivity	Reactivity	Concentration: 0.5 mg/ml
Immunogen	Immunogen	Application Note	Close: MP6-XT22
Positive control	Positive control	Suggested dilution	Immunogen: Recombinant Mouse TNF
General notes	General notes	WB	Isotype: Rat IgG1
		IHC-P	Reactivity: QC Testing: Mouse
		FACS	Storage Buffer: Aqueous buffered solution containing $\leq 0.09\%$ sodium azide
		MS	Description
		Not tested in other assays	The MP6-XT22 antibody specifically binds to mouse Tumor Necrosis Factor (TNF, also known as TNF- $\alpha$ ). TNF is produced by many activated cell types including monocytes, macrophages, astrocytes, granulocytes, mast cells, T and B lymphocytes, NK cells, keratinocytes, fibroblasts, adipocytes, and certain tumor cells. Activated cells express type II transmembrane TNF glycoproteins that associate as homotrimeric complexes. After enzymatic cleavage, the extracellular regions of membrane TNF are shed as soluble homotrimers. TNF is a potent multifunctional cytokine that can exert regulatory and cytotoxic effects on a wide range of normal lymphoid and non-lymphoid cells and tumor cells. Although TNF serves as a primary mediator in protective immune responses against microbial and viral pathogens, it can also drive systemic pathophysiological responses including septic shock, cachexia and autoimmune diseases. Mouse TNF exerts its biological activities by binding and signaling through cell surface membrane Type I and Type II TNF Receptors (aka, TNFR1/CD120a and TNFR2/CD120b, respectively).
		Calculated MW	
		Form	Immunocytochemistry using a three-step staining procedure that employs Biotin Goat anti-Rat IgG secondary antibody (Cat. No. 862205) and Anti-Rat Ig HRP Detection Kit (Cat. No. 861013) (Nomarski optics, original magnification 400X). To demonstrate the specificity of Purified Rat Anti-Mouse TNF (Cat. No. 869064), the antibody was blocked by the preincubation with excess recombinant mouse TNF (Cat. No. 864659; data not shown).
		Buffer	Preparation and Storage
		Storage	Store undiluted at 4°C. The monoclonal antibody was purified from tissue culture supernatant or ascites by affinity chromatography.
		Concentration	Application Notes
		Immunogen	Applications
		Purification	Intracellular staining (flow cytometry) Routinely Tested
		Conjugation	Immunocytochemistry Tested During Development
			Recommended Assay Procedure:
			Immunocytochemistry: The ICC format of the purified MP6-XT22 (Cat. No. 559064) antibody can be used to identify and enumerate TNF producing cells by immunocytochemistry. For optimal indirect immunocytochemical staining, the MP6XT22 antibody should be titrated ( $\leq 1 \mu\text{g}$ ) and visualized via a three-step staining procedure using Biotin Goat Anti-Rat IgG and streptavidin horseradish peroxidase (HRP). A detailed protocol for the procedure is found below. For optimal detection of cytokine producing cells, horseradish peroxidase as the preferred enzyme system.
			CYTOKINE IMMUNOCYTOCHEMISTRY PROTOCOL
			REAGENTS REQUIRED
			1. Fixation Buffer: BD Pharmingen™ ICC Fixation Buffer (BD Cat. No. 550010) or 5% formalin (10% formalin, CMS, Cat. No. 245-684) is dissolved in phosphate buffered-saline (PBS) (Bacto FA Buffer, Difco Laboratories, Cat. No. 2314-15-0)



# FAIR @source, @scale - stakeholder engagement

**74 vendors  
900 + Antibodies**

(34)	(33)	(32)	(31)	(30)	(29)	(28)
(27)	(26)	(25)	(24)	(23)	(22)	(21)
(20)	(19)	(18)	(17)	(16)	(15)	(14)
(13)	(12)	(11)	(10)	(9)	(8)	(7)
(6)	(6)	(5)	(5)	(4)	(4)	(3)
(3)	(2)	(2)	(1)	(1)	(1)	(31)
(30)	(29)	(28)	(27)	(26)	(25)	(24)
(23)	(22)	(21)	(20)	(19)	(18)	(17)



# FAIR @source, @scale – stakeholder engagement

- Fact extraction = full natural language processing (morphological analysis, part of speech analysis /semantic role labelling, NER) + zoning (for table handling etc) +rules evaluation.
- Can we use our existing models to extract contextual facts?

Product name	Anti-ACAT1 antibody [EPR10359]
Description	Rabbit monoclonal [EPR10359] to ACAT1
Host species	<b>Rabbit</b>
Tested applications	Suitable for: WB, IP, IHC-P Unsuitable for: Flow Cyt or ICC/IF
Species reactivity	Reacts with: <b>Human</b> Predicted to work with: <b>Mouse, Rat</b>
Immunogen	Synthetic peptide corresponding to <b>Human</b> AT1 aa 350-450. Database link: P24752
Positive control	Human fetal liver, SW480, Jurkat, HepG2, (C) SKBR-3, THP-1 and Human fetal kidney lysates, Human heart & liver tissue, immunoprecipitation pellet from Human fetal liver lysate
General notes	<p>This product is a recombinant monoclonal antibody, which offers several advantages including:</p> <ul style="list-style-type: none"> <li>- High batch-to-batch consistency and reproducibility</li> <li>- Improved sensitivity and specificity</li> <li>- Long-term security of supply</li> <li>- Animal-free production</li> </ul> <p>For more information <a href="#">see here</a>.</p> <p>Our RabMAb® technology is a patented hybridoma-based technology for making rabbit monoclonal antibodies. For details on our patents, please refer to <a href="#">RabMAb® patents</a>.</p> <p>The Life Science industry has been in the grips of a reproducibility crisis for a number of years. Abcam is leading the way in addressing the problem with our range of recombinant monoclonal antibodies and knockout edited cell lines for gold-standard validation.</p> <p>One factor contributing to the crisis is the use of antibodies that are not suitable. This can lead to misleading results and the use of incorrect data informing project assumptions and direction. To help address this challenge, we have introduced an application and species grid on our primary antibody datasheets to make it easy to simplify identification of the right antibody for your needs.</p>



# Modelling the required Facts

- Identify the **facts** to be extracted
- Identify the **context** that contains the facts
- Create a document **fingerprint** that identifies the document type using appropriate context
- Create **extractors** to extract the facts using specialized context and fact types

The screenshot shows a software interface for fact extraction. The interface is in English (en) and has a search bar at the top. Below the search bar, there are several sections:

- CONTENT TYPES**: A list of content types, including Antibody document, Catalog number fact, Clonality fact, Clone fact, Concentration fact, Conjugate facts, Host fact, and Manufacturer fact.
- FACT EXTRACTORS**: A list of fact extractors, including Manufacturer from taxonomy anywhere (highlighted in blue), Manufacturer from website near document start, Manufacturer near document start, Name facts, Storage fact, Type facts, and Unreadable document.
- FACT NAMES**: A list of fact names.
- FRAMEWORK SETTINGS**: A list of framework settings, including Clonality, Conjugate, Manufacturers, Manufacturer taxonomy, Species, and Storage.

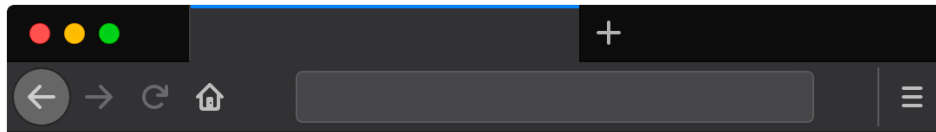
Red arrows point from text labels to specific elements in the interface:

- "Identify content to be classified" points to the "Antibody document" content type.
- "Extractors that extract facts" points to the "Manufacturer from taxonomy anywhere" extractor.
- "Identify facts to be extracted" points to the "FACT NAMES" section.
- "Identify grammatical units and structures" points to the "FRAMEWORK SETTINGS" section.
- "Extraction using models" points to the "Manufacturer taxonomy" framework setting.





# FAIR @source, @scale – stakeholder engagement

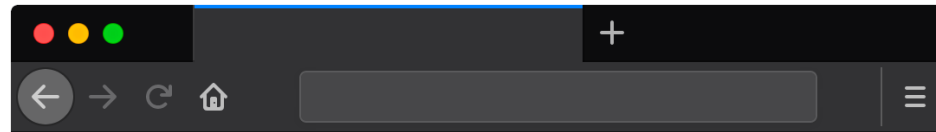


## Antibody datasheets classification

How to use: Type abcam id e.g. ab14715 or ab46545

Antibody id:

[Get Antibody details](#)



## Antibody datasheets classification

How to use: Type abcam id e.g. ab14715 or ab46545

Antibody id:

[Get Antibody details](#)

**Clonality:** Polyclonal

**Conjugate:** HNE conjugated to BC.

**Host:** Rabbit

**Manufacturer:** Abcam

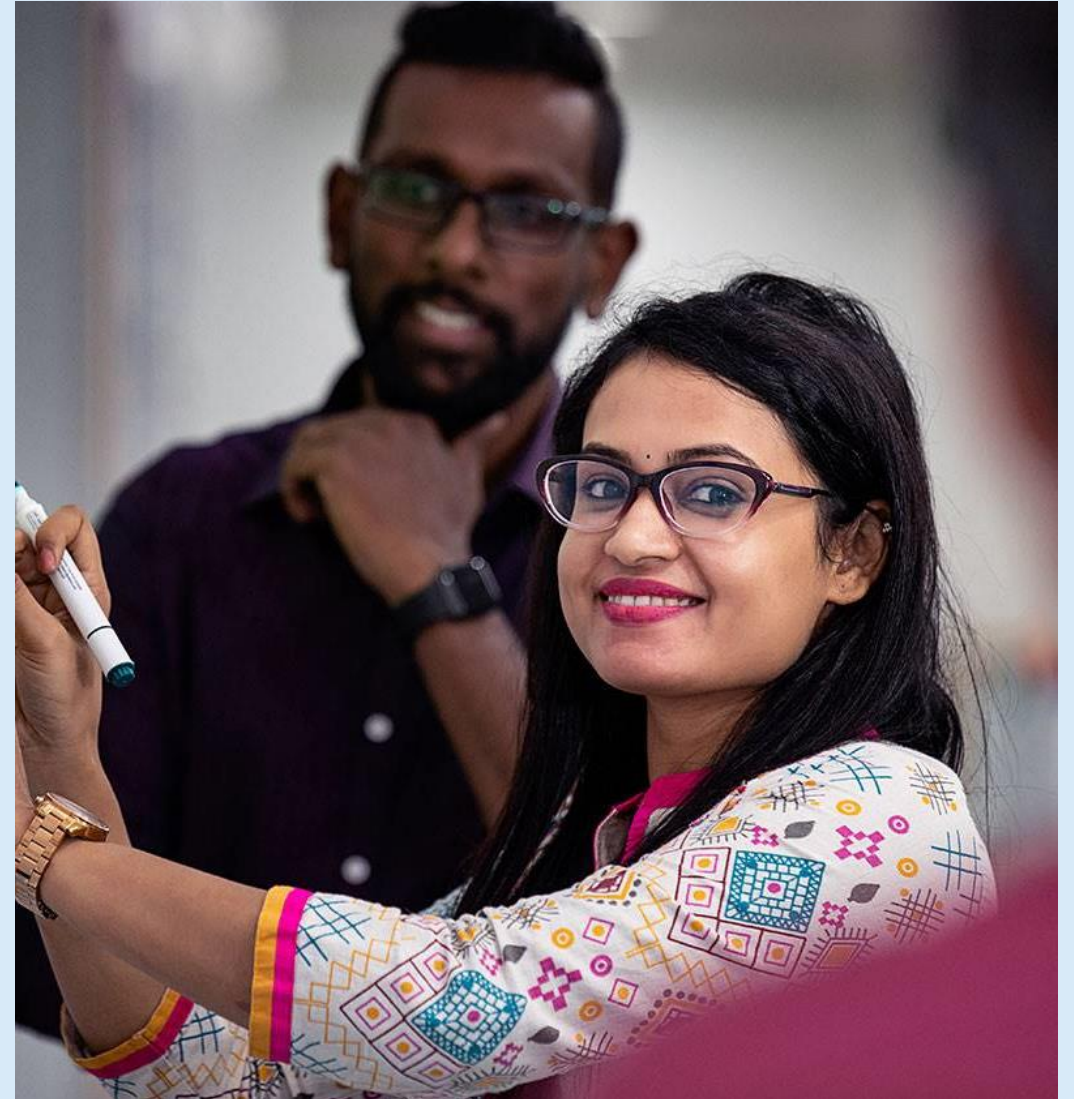
**Name:** Anti-4 Hydroxynonenal antibody ab46545

**Storage:** Shipped at 4°C. Store at +4°C short term (1-2 weeks). Upon delivery



# Change Management - Data Stewards

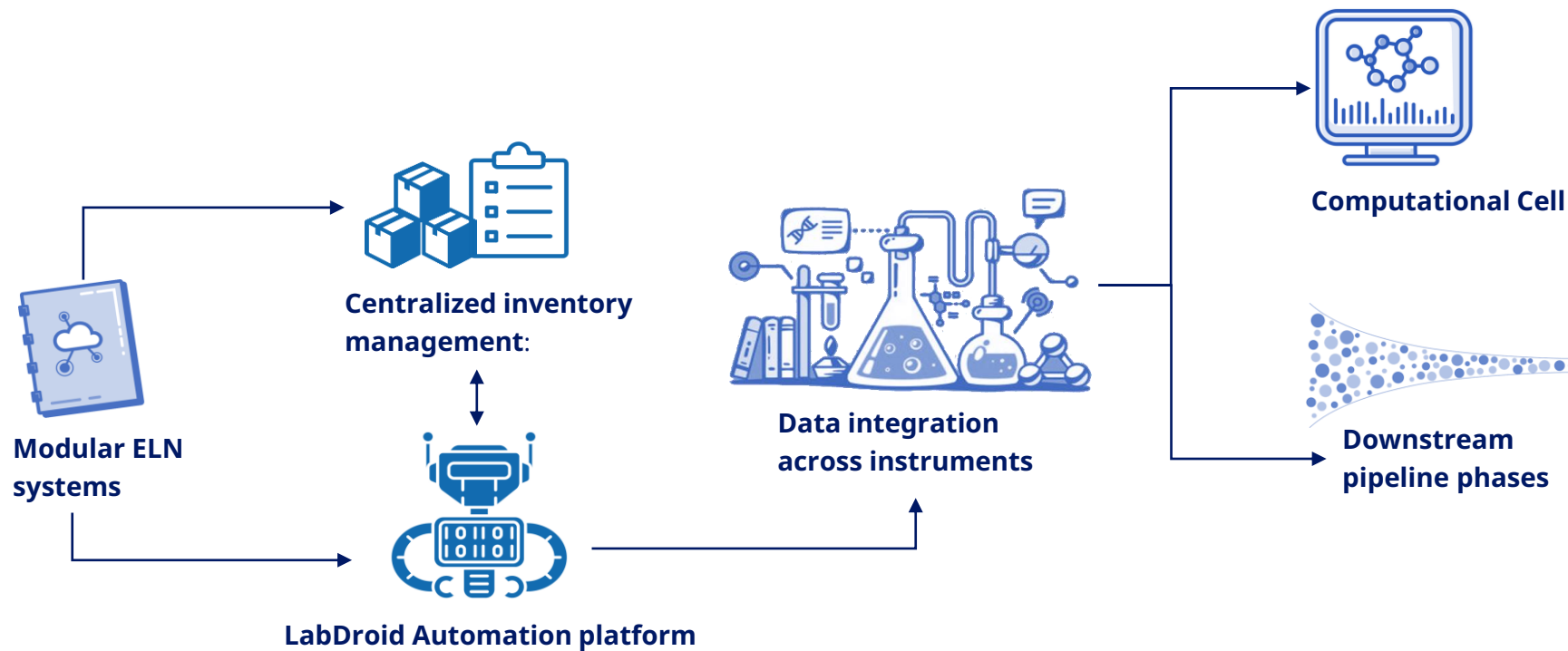
- New role of a data steward to enforce data consistency
- Previously scientists generates data and distributes
- Now the data steward can take over the governance and provide consulting expertise on projects
- No longer “my lab, my data” but data is a shared asset for everyone



# FAIR infrastructure

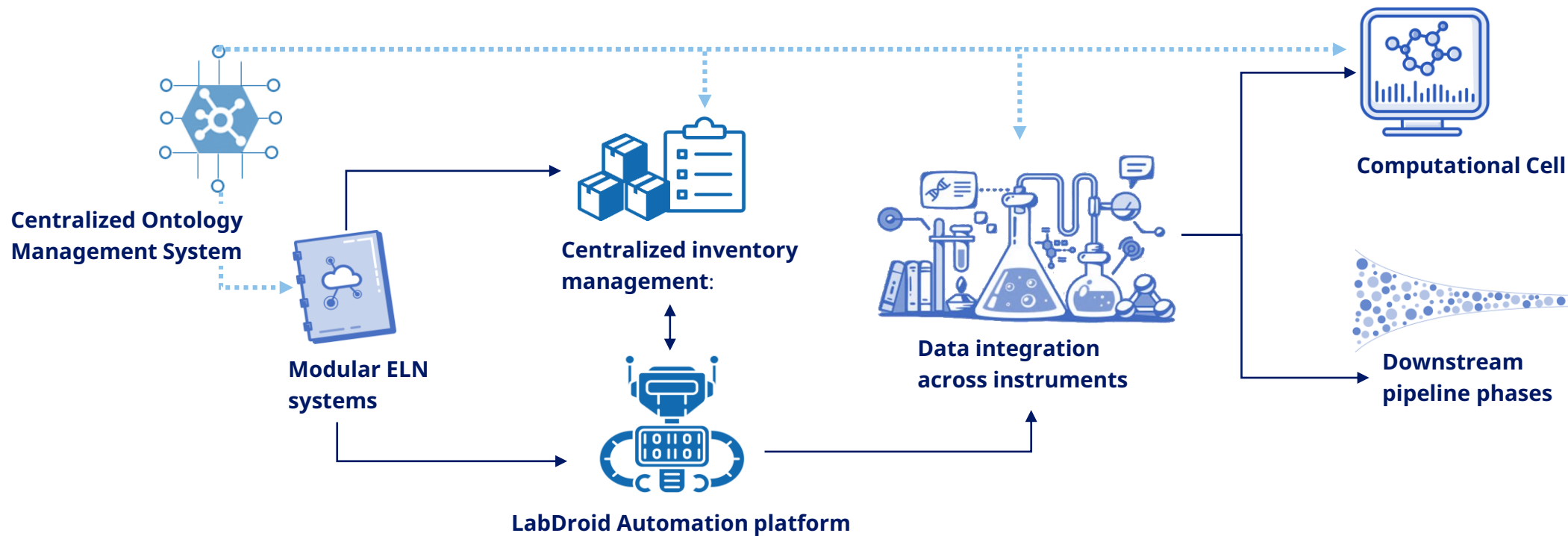


# Implement : FAIR at source



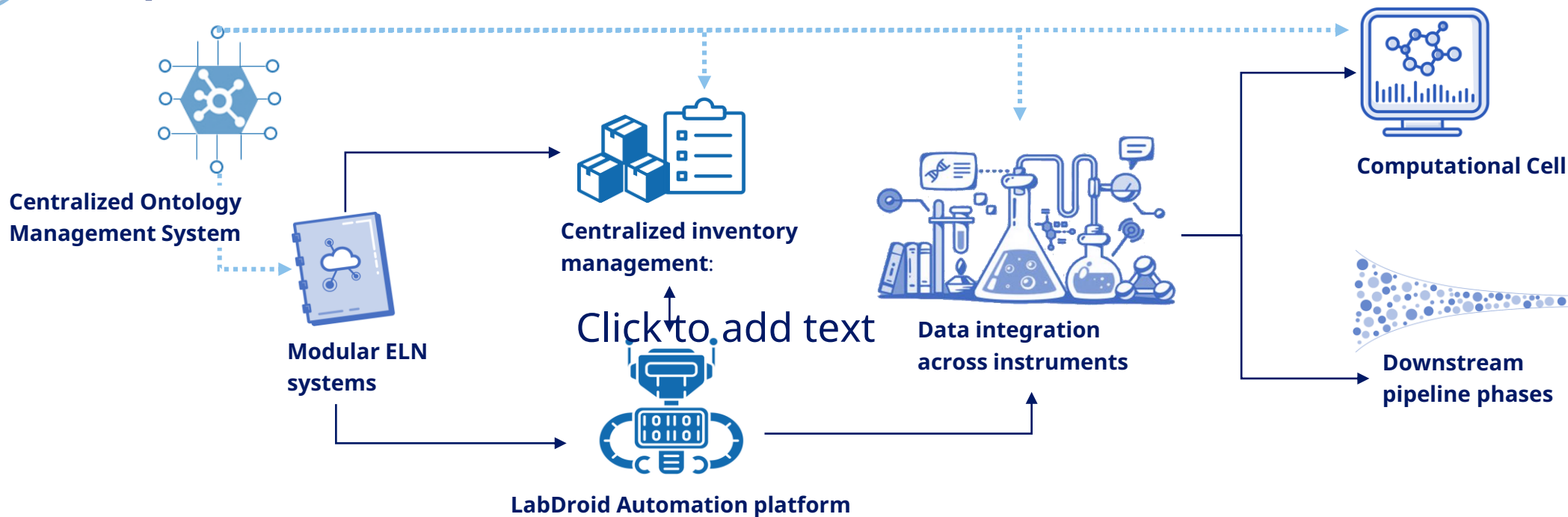


# Implement : FAIR at source





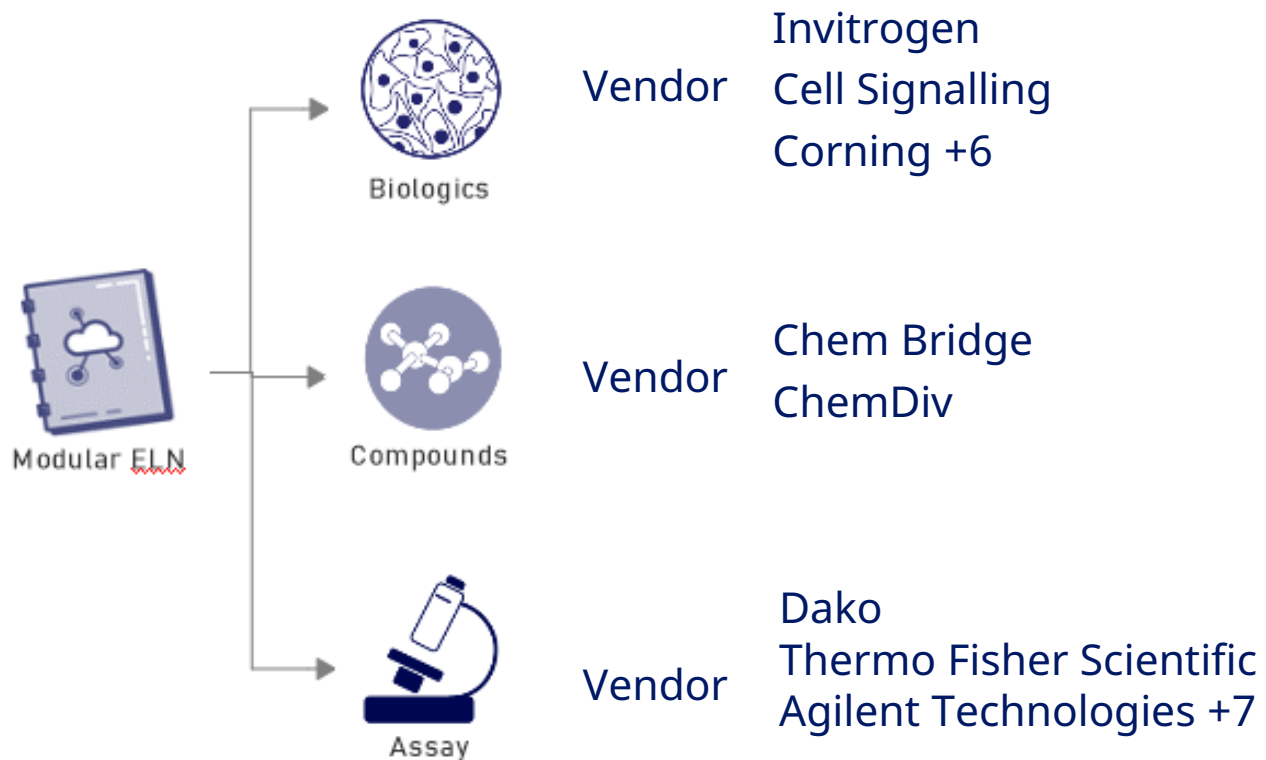
# Implement : FAIR at source



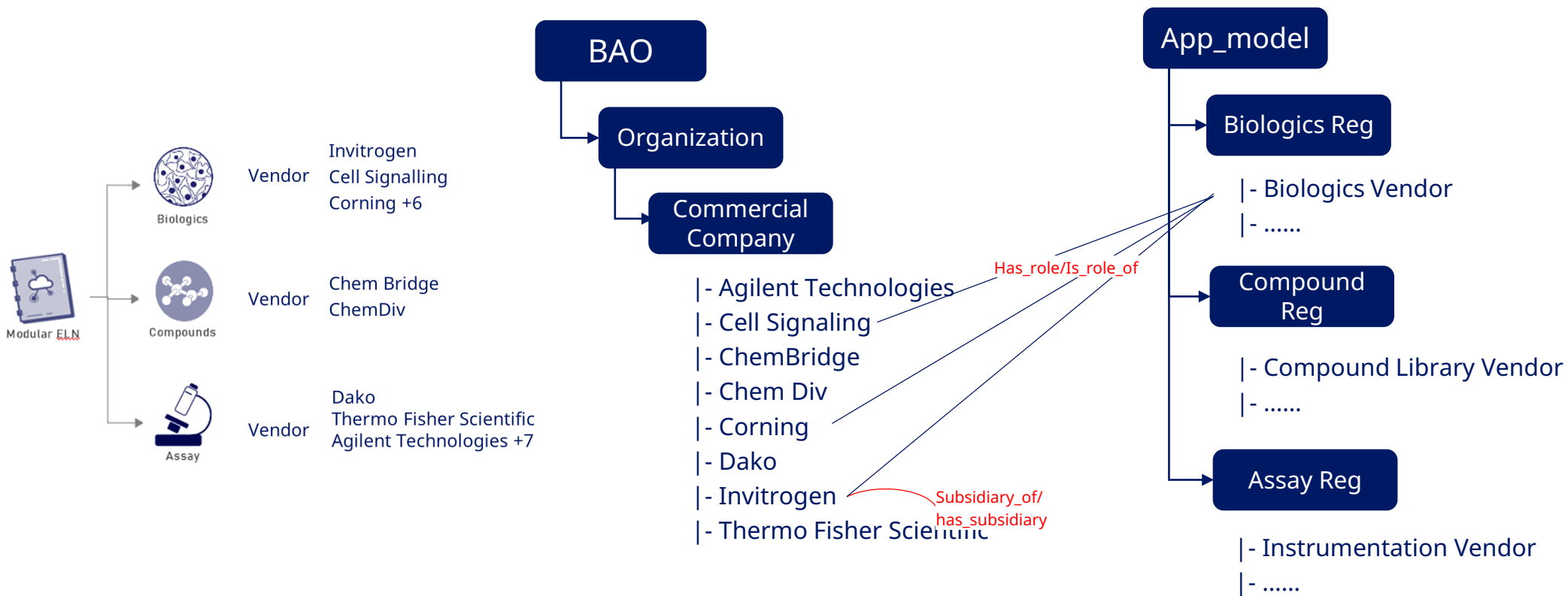
JSON terms		Mapped ontology term	
injection: {		<b>Term ID</b>	<b>Preferred label</b>
id: "2309",		http://purl.allotrope.org/ontologies/process#AFP_0001611	injection (chromatography)
time: {			
acquisition: "2019-02-06T07:14:04.000+00:00"		http://purl.allotrope.org/ontologies/result#AFR_0001158	acquisition time
},			
run_time: {		http://purl.allotrope.org/ontologies/result#AFR_0000951	duration
value: 8,	value: 8,	http://purl.obolibrary.org/obo/UO_0000031	values are not ontologized
unit: "Minutes"	unit: "Minutes"		minute
},			
volume: {		http://purl.allotrope.org/ontologies/result#AFR_0001577	injection volume setting
value: 8,	value: 8,		

Semantic interoperability through centralized vocabulary management system

# Data is the driver for governance not applications



# Data is the driver for governance not applications

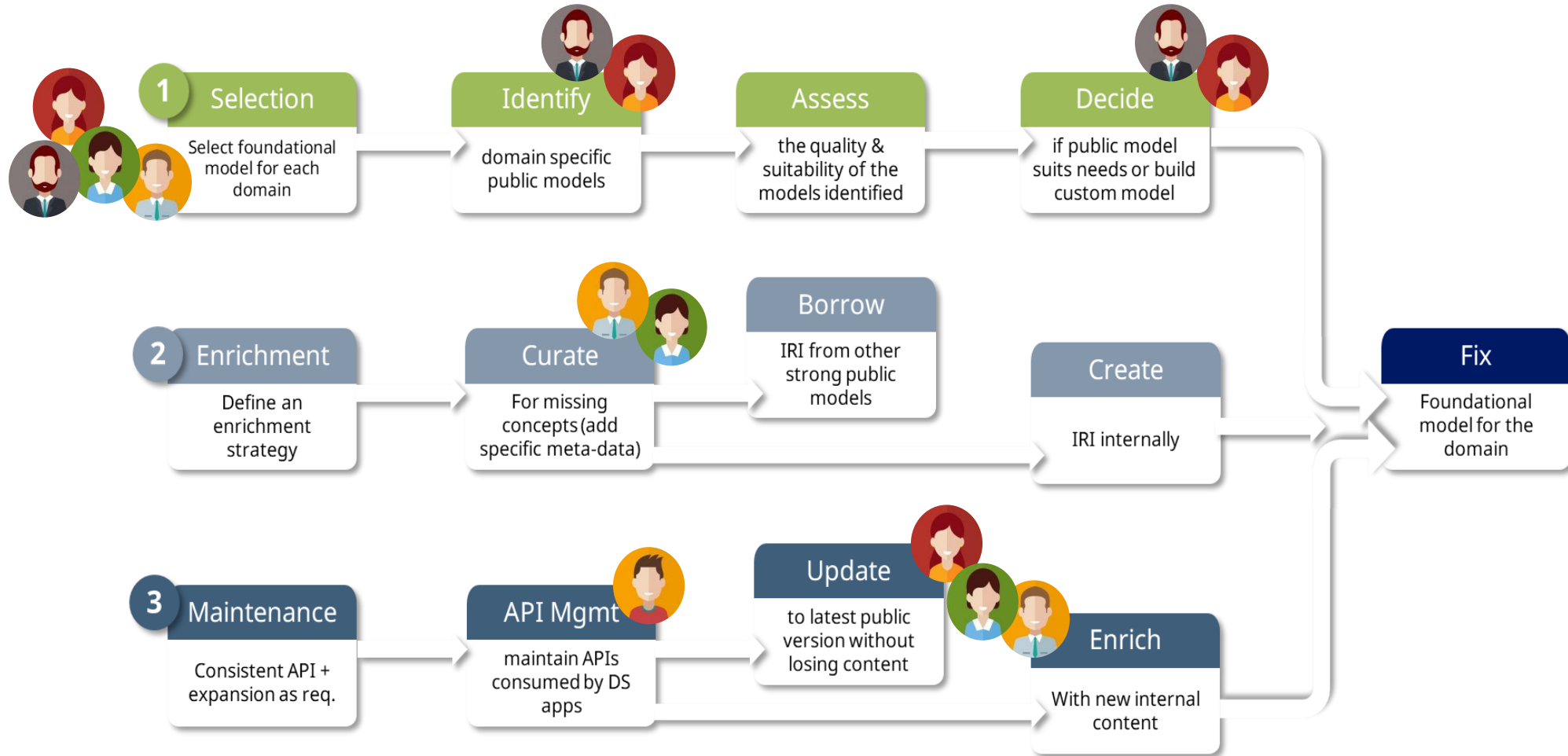


API strategy: show **Commercial Company** that are linked to **Biologic Reg** by relation **"Is role of"**





# FAIR @source, @scale – stakeholder engagement



# It is all about people



Create a digital mindset – internal & external trainings



Hire and retain talent – a global workplace



Incentivize FAIR efforts across the organization



ново nordisk®

# “Good” Data is a strategic asset

