Symposium

Advancements in Metadata and HDA Data Catalog: Exploring Cutting-Edge Developments

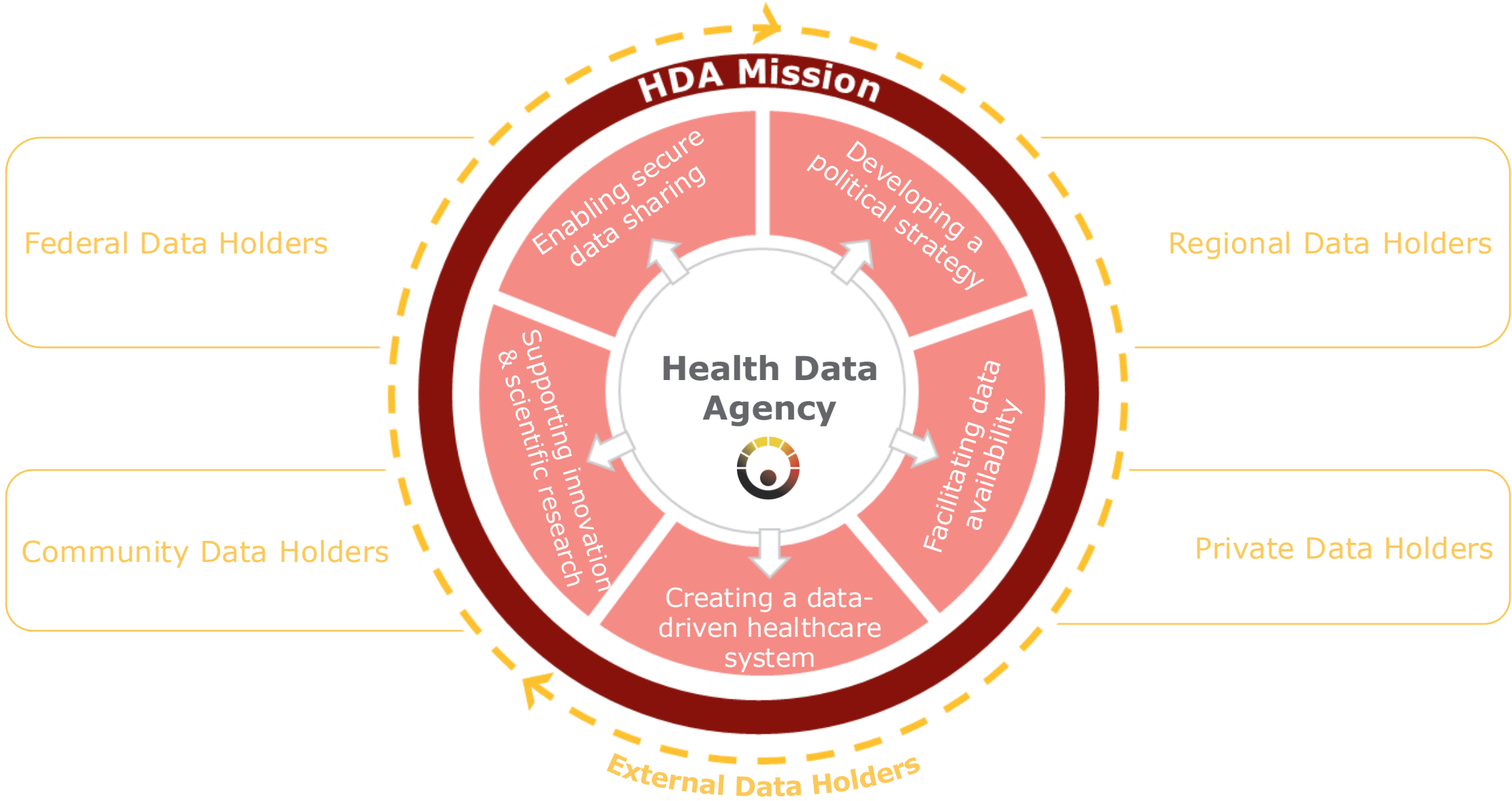
HDA

Agenda

1	Context
2	Data catalogue fundamentals
3	What’s next
4	Conclusion
5	Q&A

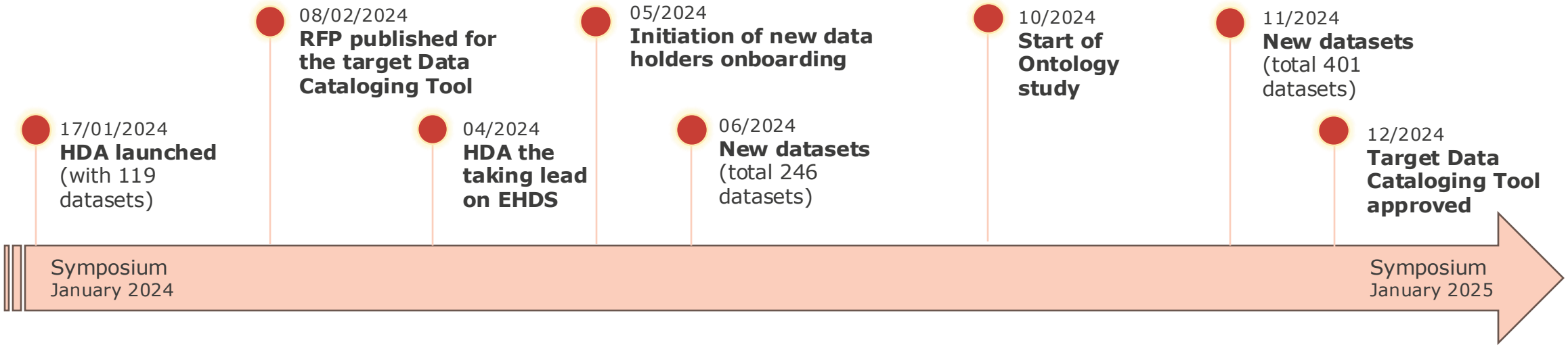
Context

HDA’s mission



HDA

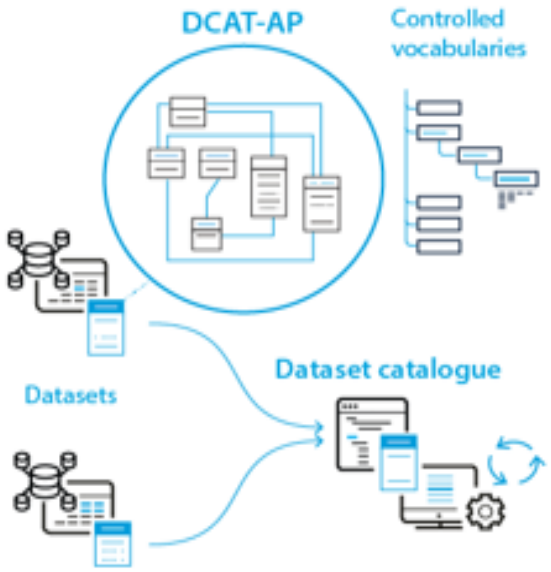
What happened in 2024?



EHDS

- *European Health Data Space (EHDS)*
- Provides a common basis for standardized sharing of health data within Europe.
- Designed to improve interoperability between health systems and data repositories in Europe, making it easier to exchange data and exchange health data across borders and domains for healthcare, research and innovation purposes.
- Metadata documentation based on DCAT-AP with DCAT vocabulary, mandatory, recommended, and optional metadata elements customized for the European context.
- Developed and maintained by the European Commission it is a cornerstone of the EU's Digital Health Strategy, fostering a more Interoperable Europe in the health sector.

DCAT-AP allows local datasets to be linked to parent catalogue(s). The use of controlled vocabularies allows different datasets to be linked regardless of the platform used.



DCAT-AP

- *DCAT Application Profile (DCAT-AP)*
- Provides a common basis for standardized description of metadata and datasets within Europe.
- Designed to improve interoperability between European data portals, making it easier to exchange data across borders and domains.
- EHDS is based on established frameworks, such as DCAT-AP and other EU data standards, with added domain-specific extensions and requirements tailored to health data governance.
- Developed and maintained by the European Commission as part of the SEMIC action for an Interoperable Europe.

Data Catalogue fundamentals

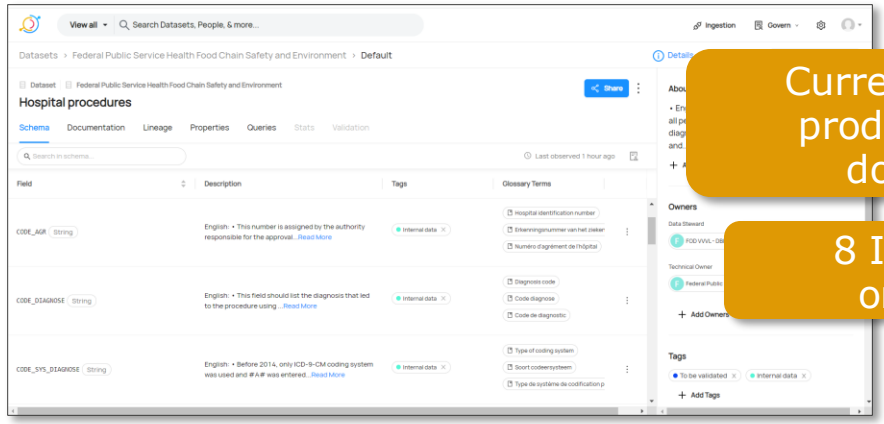
Fundamentals of the data catalogue

Facilitate data access and discovery

Inventory of data

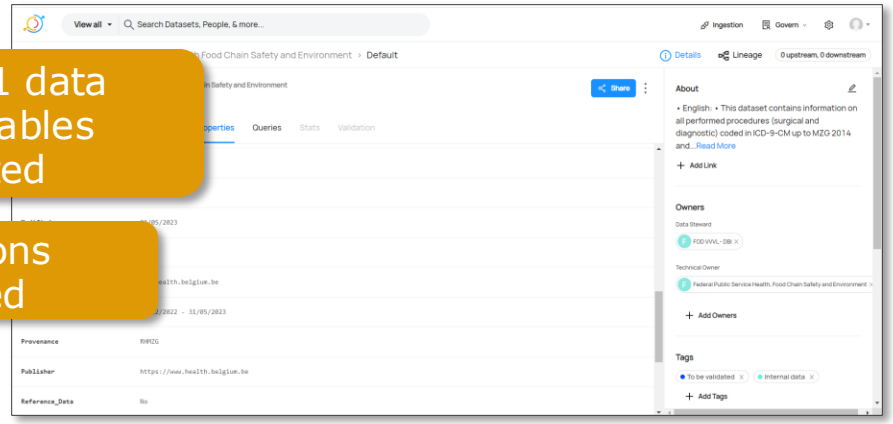
Documentation of data

Technology

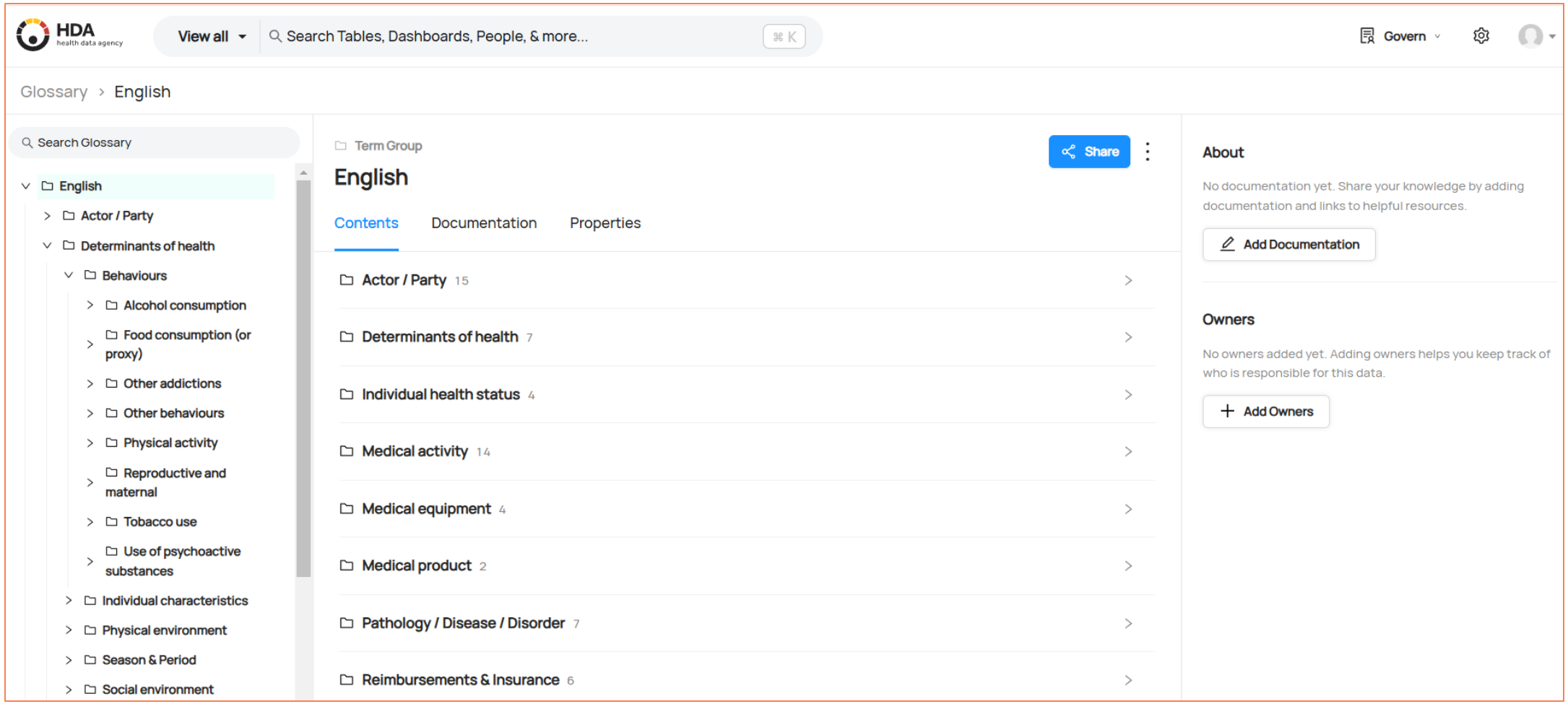


Currently 401 data products & tables documented

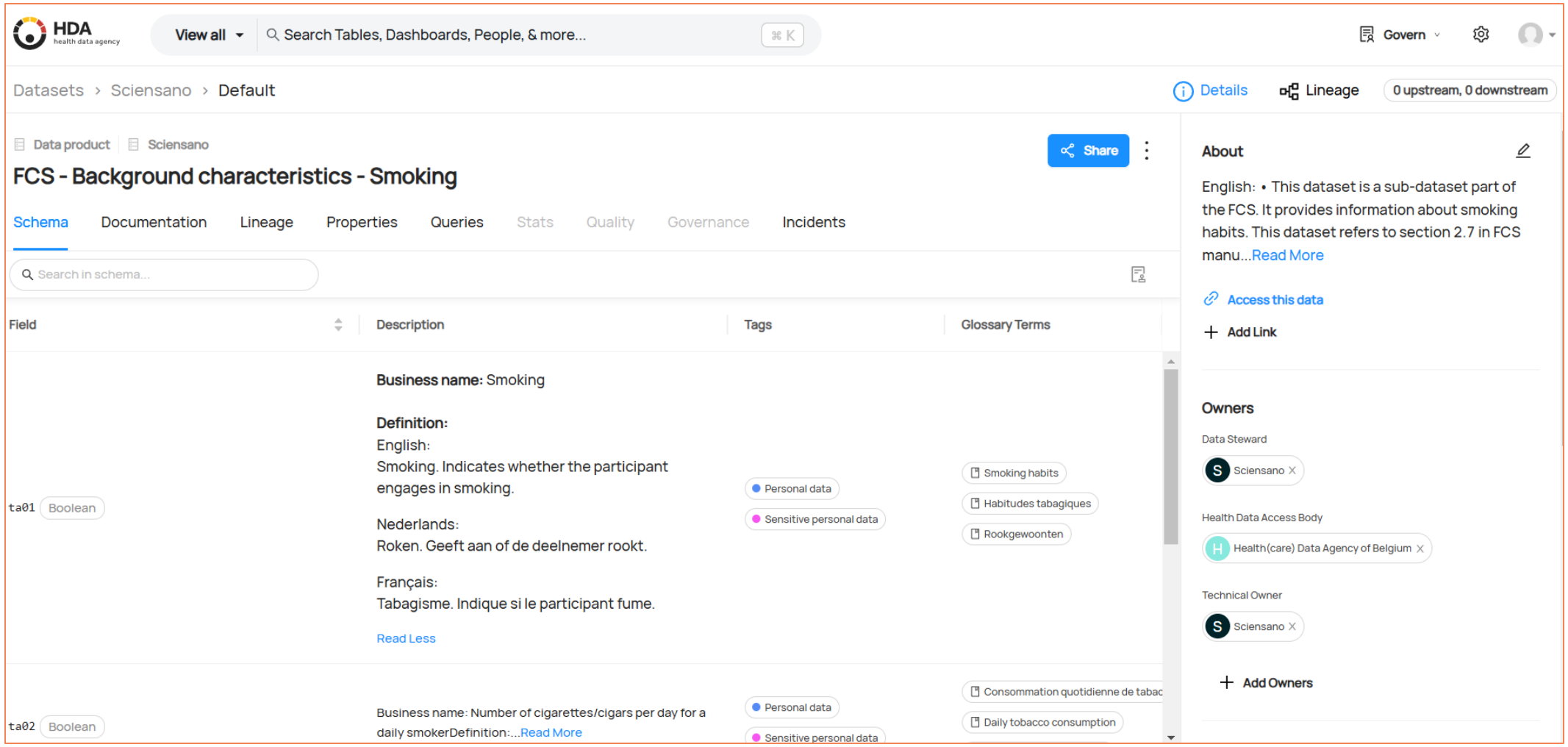
8 Institutions onboarded



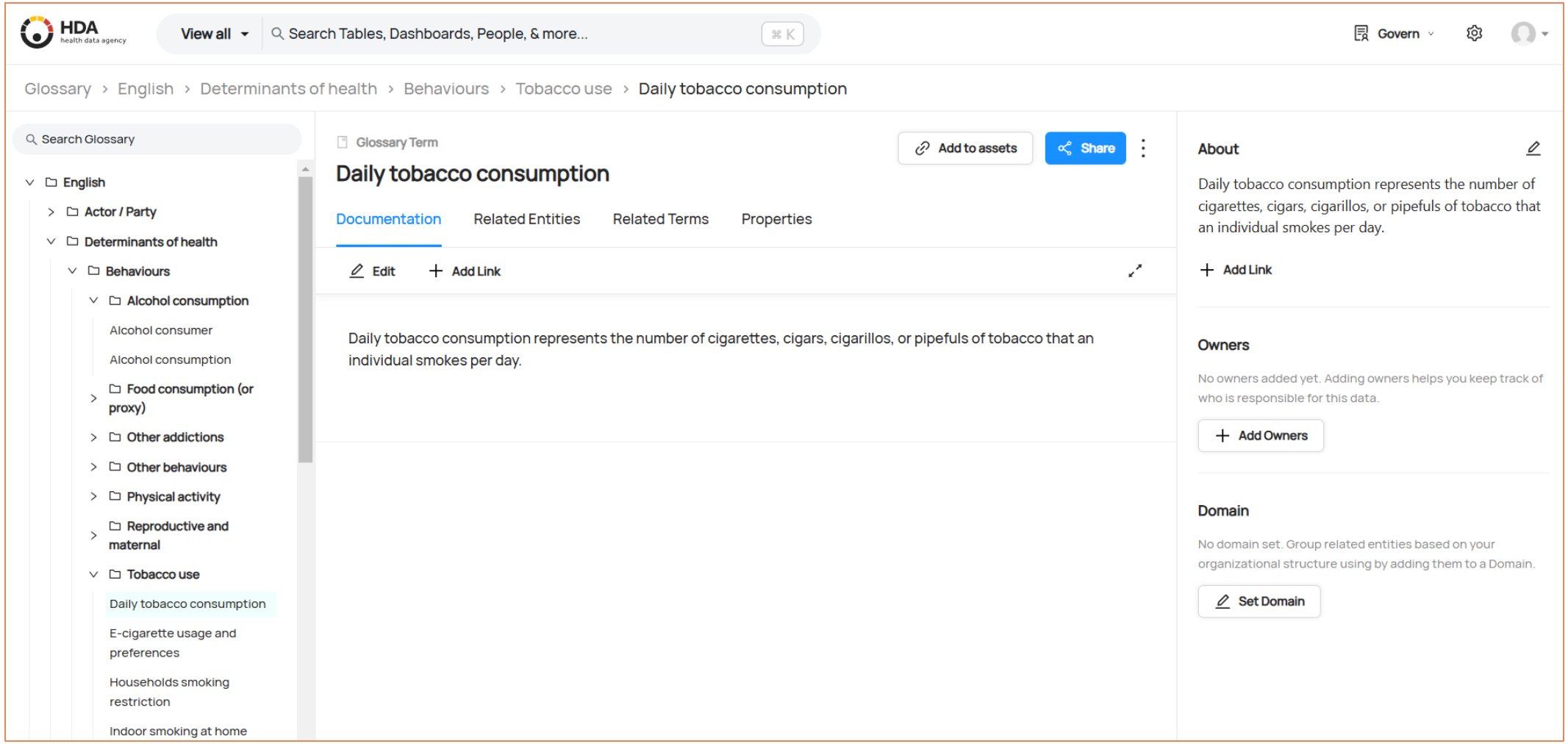
Common data model



Data dictionary



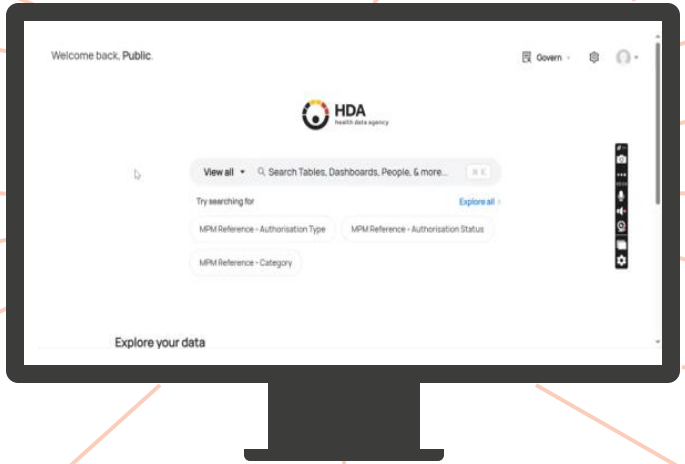
Business glossary



What does the catalogue cover

Common Data Model coverage: **66%**

- Healthcare reimbursement
- Actor/Party
- Genomics
- Socio-Economic Data
- Medical Products Authorisations
- Determinants of health & health status
- Food Consumption
- Diseases Registries
- Medicinal product reference tables
- COVID-19
- Hospital Activities
- Medical Equipment
- Cancer
- Professional reintegration
- Hospital Financials
- Other Pathologies/Diseases/Disorders



HDA

What’s next?

Improving the user experience

In order to **kickstart the HDA** initiative as swiftly as possible, we had to begin with a **fairly basic catalogue environment**. This naturally resulted in a somewhat simplistic user experience.

The goal for the upcoming months is to **enhance this experience** by introducing **new features**, also made possible by migrating to an **advanced cataloging tool**.

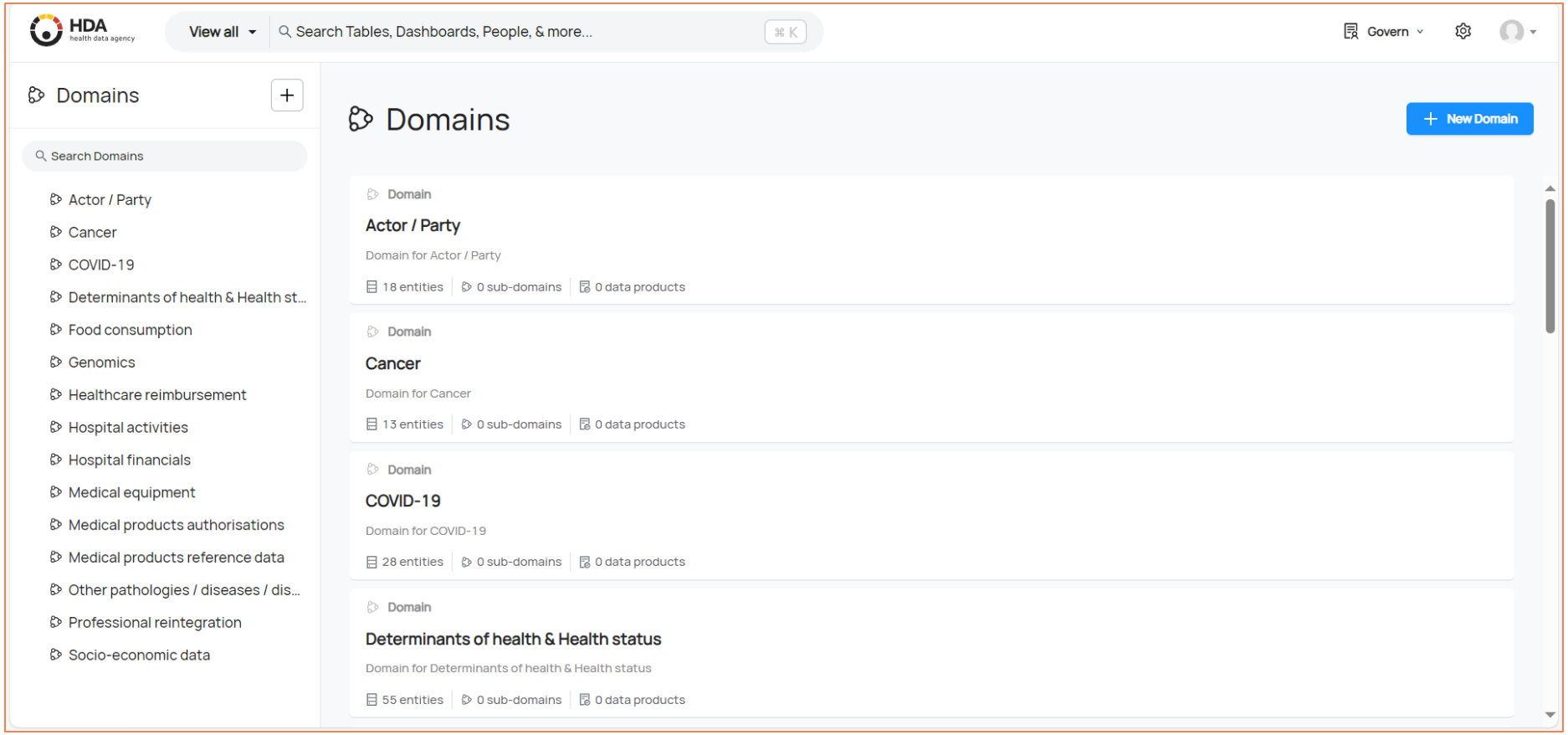
New features to come in 2025

- **Domains view:** grouping datasets per theme
- **Data models**: giving an overview of interconnected datasets
- **Reference data:** analysis of differences and documentation of reference data
- **New types of data:** data requests, data pods, data from reports, …
- **Ontology** : knowledge graph showing relationships between datasets
- **AI catalogue assistant**: Search data through a chatbot (like ChatGPT)

HDA

Domains view

Domains, another way to browse data



HDA

Data models

One table is often not enough …

I have a question regarding insights related to disability cases

Data User

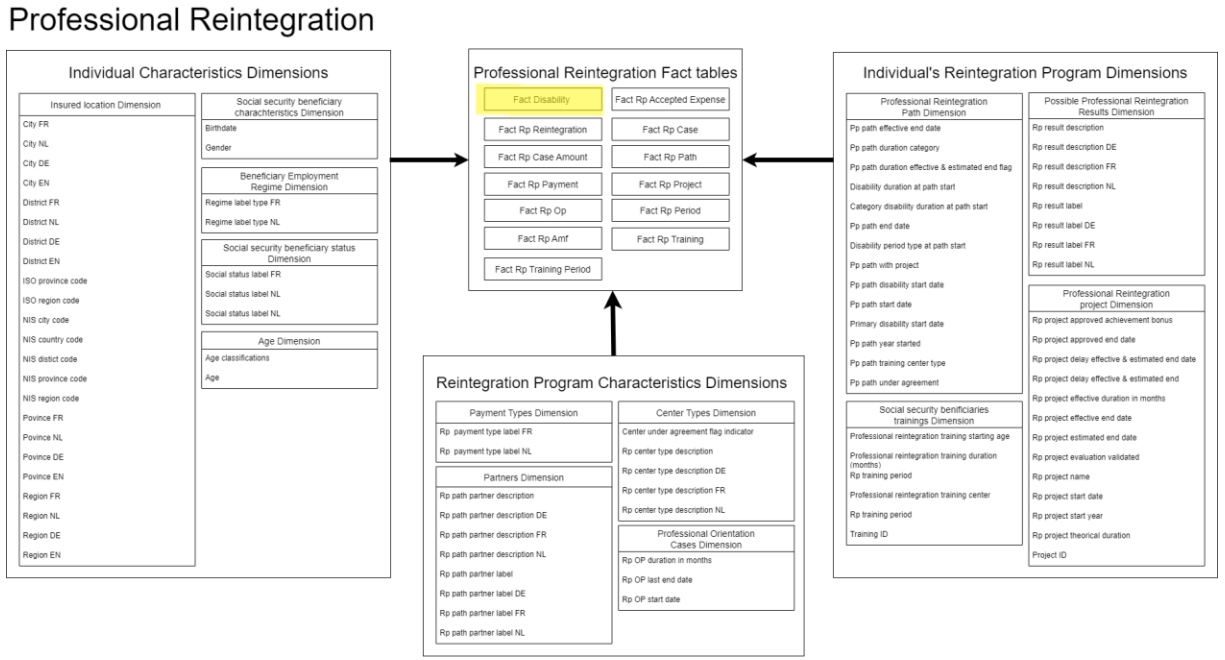
It's a challenging question because **the necessary information isn't consolidated into a single table or existing data product**. I'll need to examine various tables from different sources to provide you with an answer.

(HDA) Request handler

Data models as a solution to show relationships between datasets

(HDA) Request handler

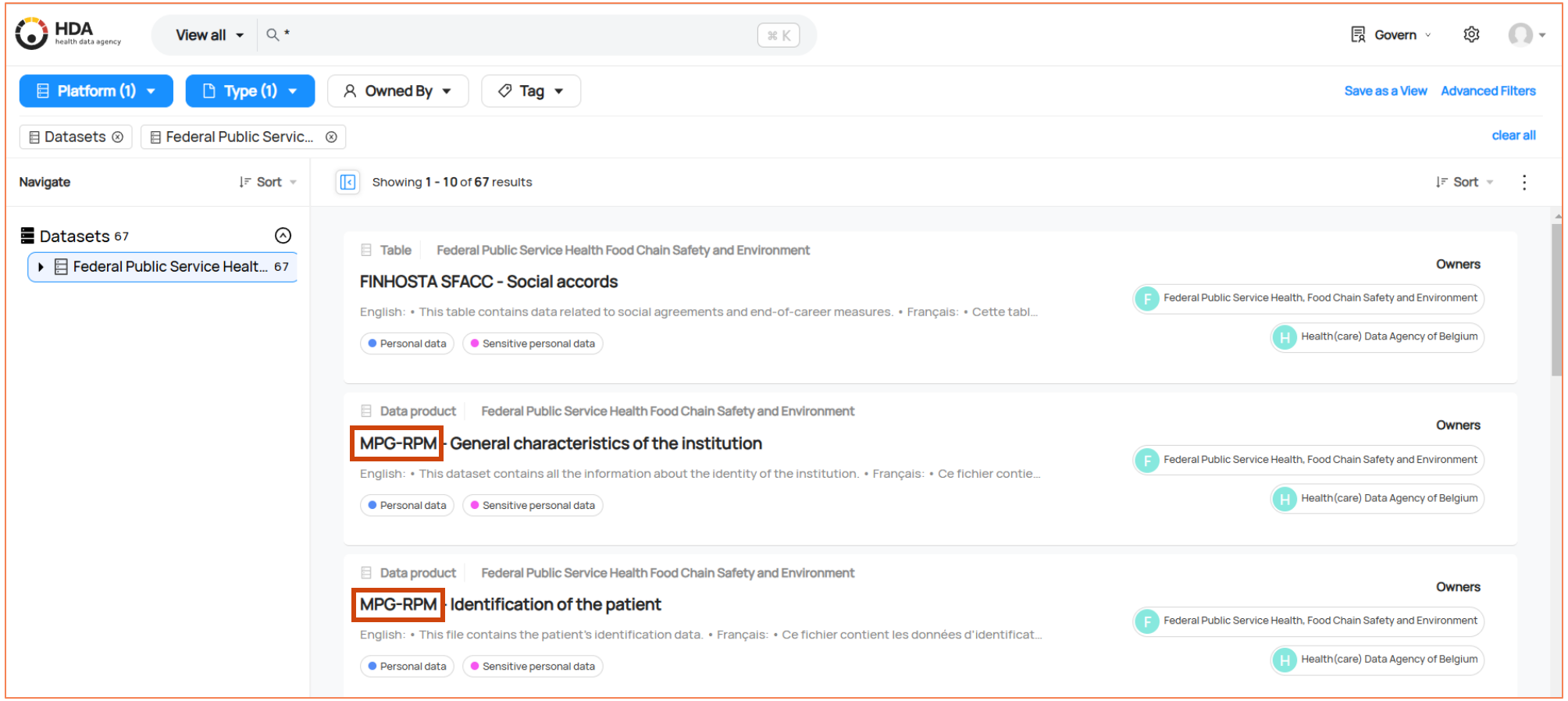
By **combining multiple tables** such as the Disability Fact Table with dimension tables Individual Characteristics, Reintegration Projects and Payment Types, you should be able to perform a **meaningful and in-depth analysis.**



Data models are now being documented in the catalogue

HDA

Groups get highlighted in titles too!



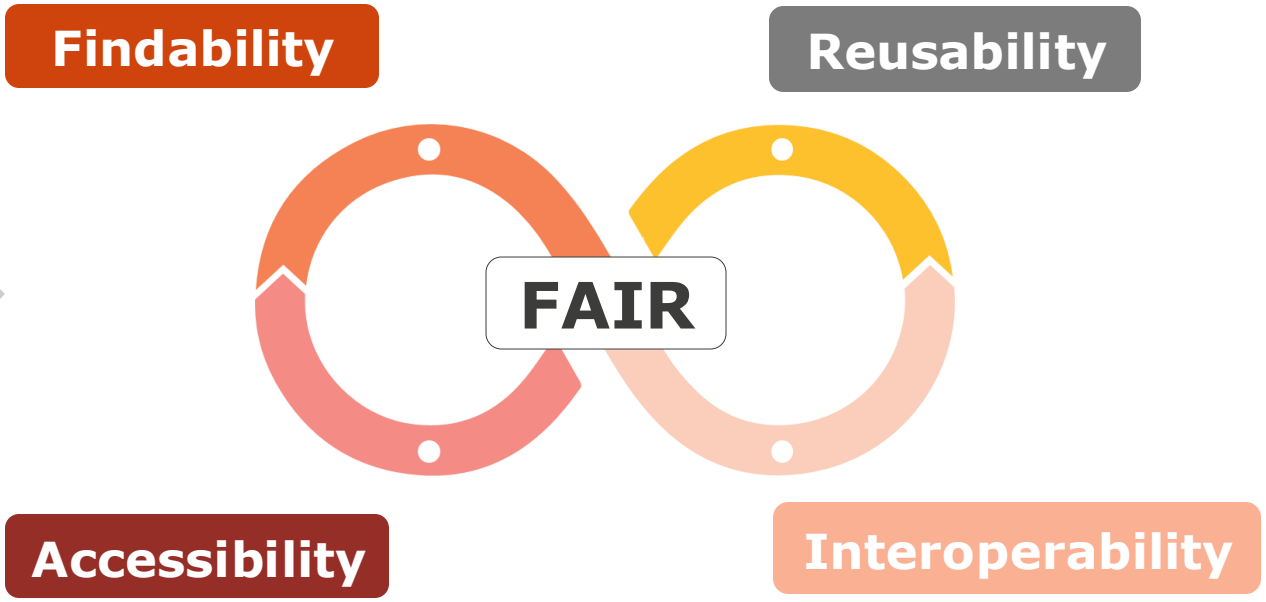
HDA

Reference data

The end goal of our reference data process is…

1. …to determine which datasets can be **combined to provide** comprehensive **answers** to (complex) data requests.

2. …to determine which **reference lists** should be **standardized** at and across data holders.



We compared values for 223 columns from 84 datasets

We identified (some) similarities for 41 cases (columns):

Less than half Match — 17 cases

More than half Match — 16 cases

Full Match — 8 cases

Example:

Dataset ID 291

Identifier value	Business Value NL
1	Landsbond der Christelijke Mutualiteiten
2	Landsbond van de Neutrale **Ziekenfondsen**
3	Nationaal Verbond van Socialistische Mutualiteiten
4	Landsbond van Liberale Mutualiteiten
5	Landsbond van de Onafhankelijke Ziekenfondsen
6	Hulpkas voor Ziekte- en Invaliditeitsverzekering
9	Kas der Geneeskundige Verzorging van de Nati

Limited differences

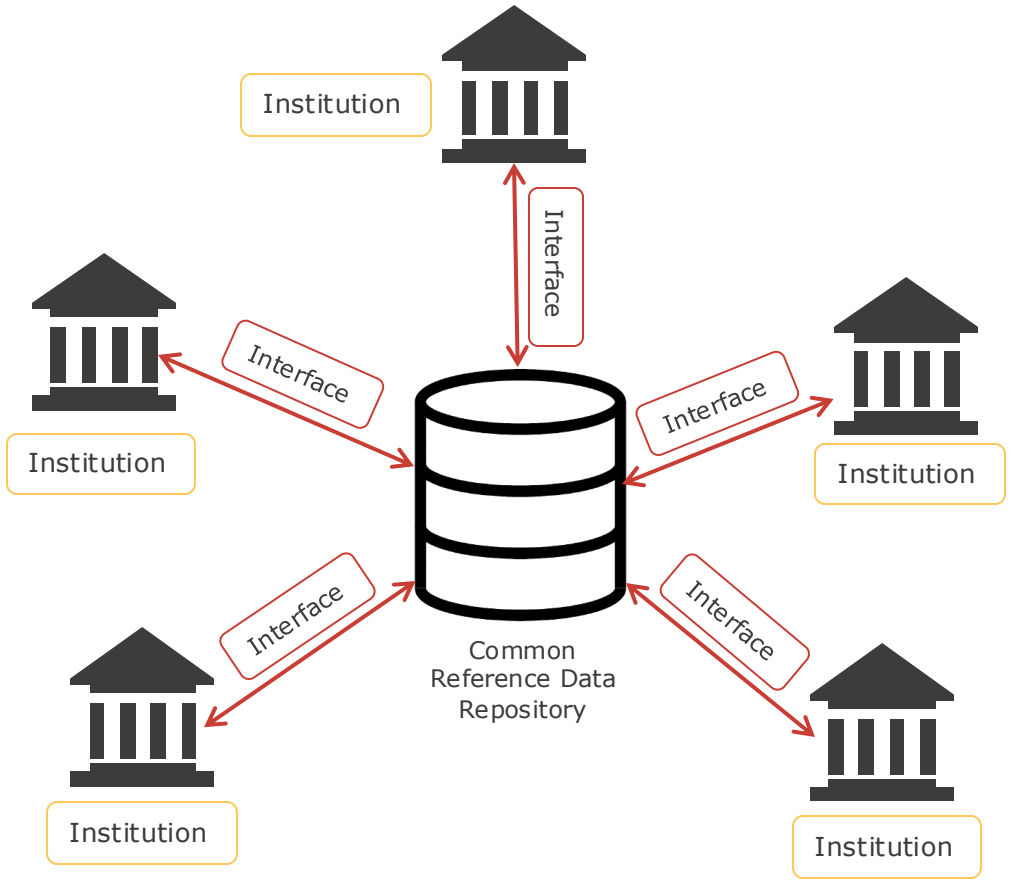
Matching Percentage 86%

Dataset ID 296

Identifier value	Business Value NL
1	Landsbond der Christelijke Mutualiteiten
2	Landsbond van de Neutrale **Mutualiteitsverbonden**
3	Nationaal Verbond van Socialistische Mutualiteiten
4	Landsbond van Liberale Mutualiteiten
5	Landsbond van de Onafhankelijke Ziekenfondsen
6	Hulpkas voor Ziekte- en Invaliditeitsverzekering
9	Kas der Geneeskundige Verzorging van de Nationale Maatschappij der Belgische Spoorwegen

HDA

Towards a common reference data repository



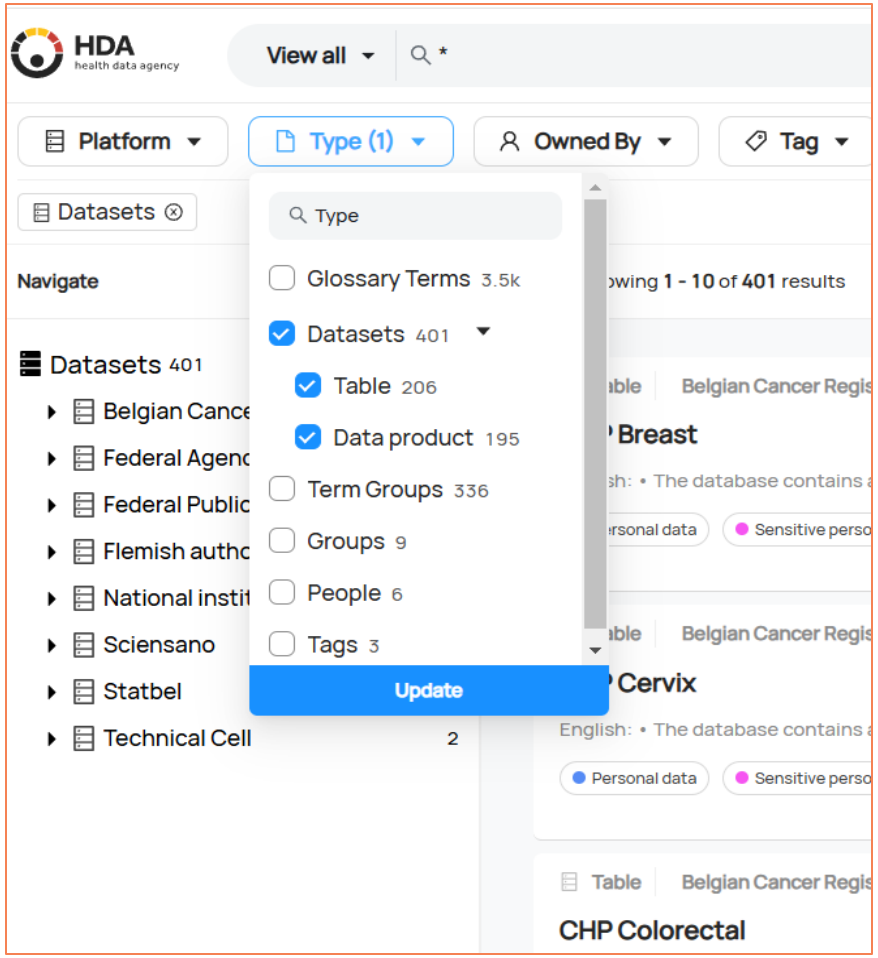
The advantages:

- **Clear Ownership**: Enables focused maintenance responsibilities, reducing duplication of effort.
- **Standardization:** Provides consistent set of data to ensure all systems and applications use the same values.
- **Governance and Control:** Includes policies and procedures to manage changes to the reference data.
- **Interoperability:** Enables systems to communicate effectively by referencing a unified set of data.
- **Accessibility:** Makes reference data easily accessible to authorized users and systems.

Reference data will be added to the catalogue

New types of data

Expanding the catalogue further



We are considering adding new types of data, such as:

- Data from reports (e.g. studies)
- Data requests
- Data from data pods
- …

Ontology

Understanding each other's is key

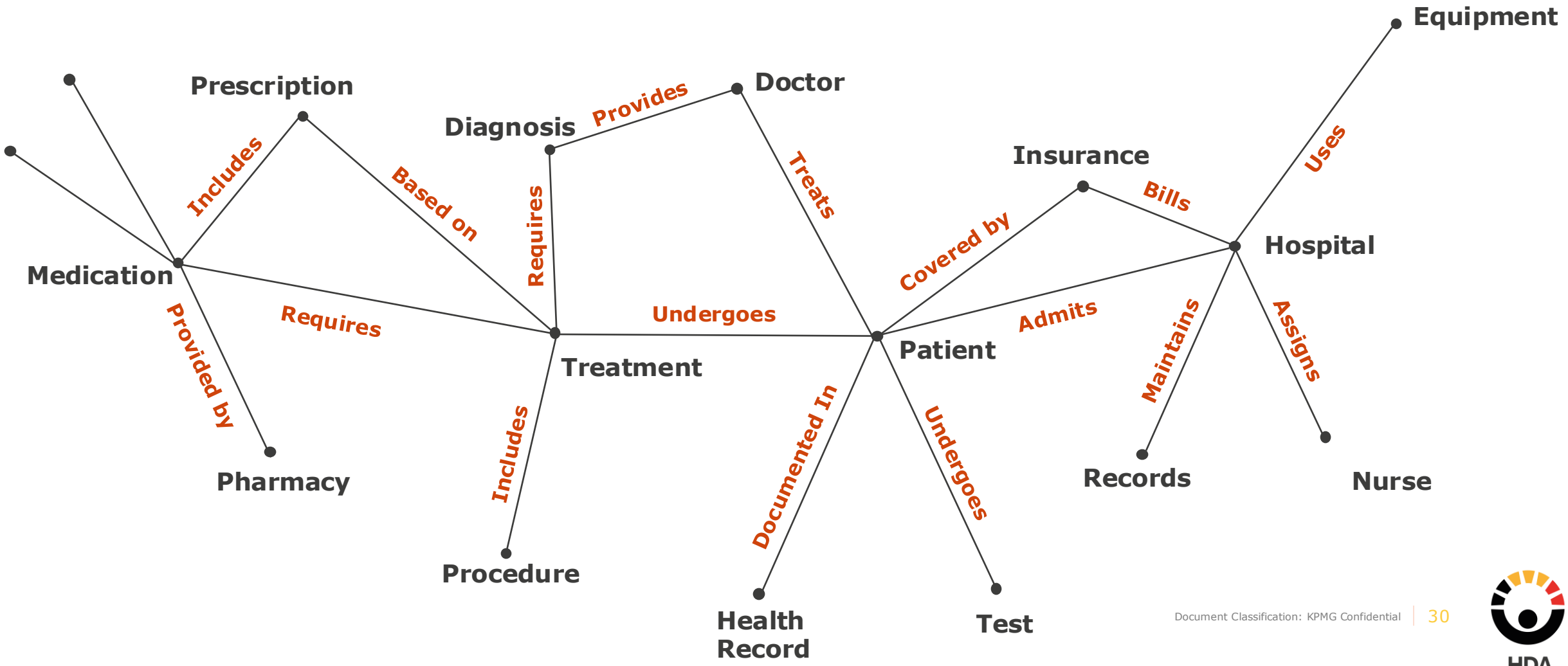
Data User

End users struggle to accurately interpret and integrate data from multiple sources due to **inconsistent terminology**, **ambiguous relationships**, and a **lack of shared understanding** of key domain concepts. This leads to confusion and inefficiencies when analyzing data and deriving insights.

An **ontology** is needed to establish a shared understanding of key domain concepts, standardize terminology, and define clear relationships between data elements. By implementing an ontology, data users can consistently interpret and integrate data across multiple sources, enabling accurate analysis and efficient decision-making.

(HDA) Request handler

Ontologies help improving the understanding of the health data environment



HDA

AI catalogue assistant

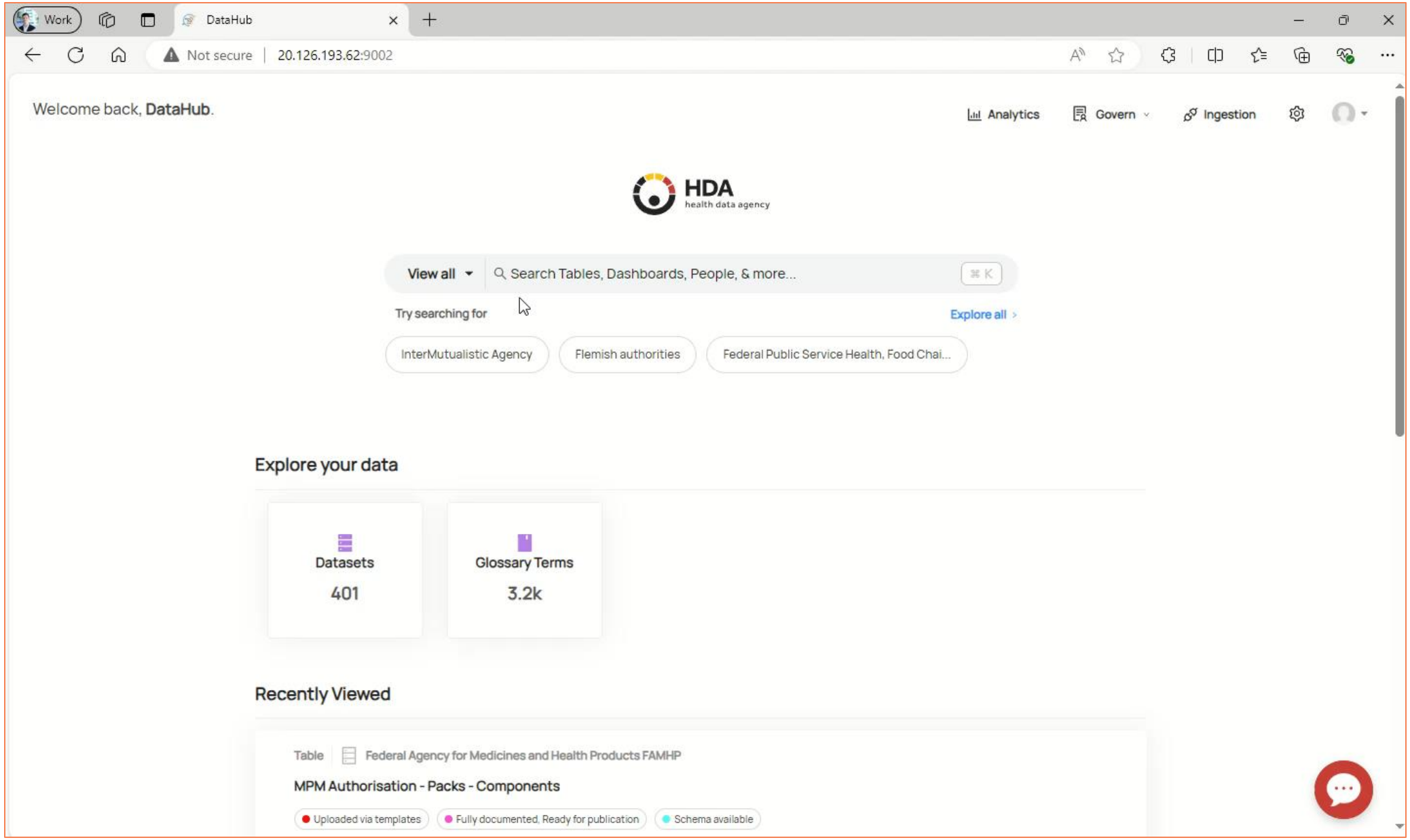
Sneak peek of HDA's new AI catalogue assistant

End users want to leverage the data catalogue for retrieving relevant datasets…

…But even in a data catalogue, it is not always easy to use the right search terms…

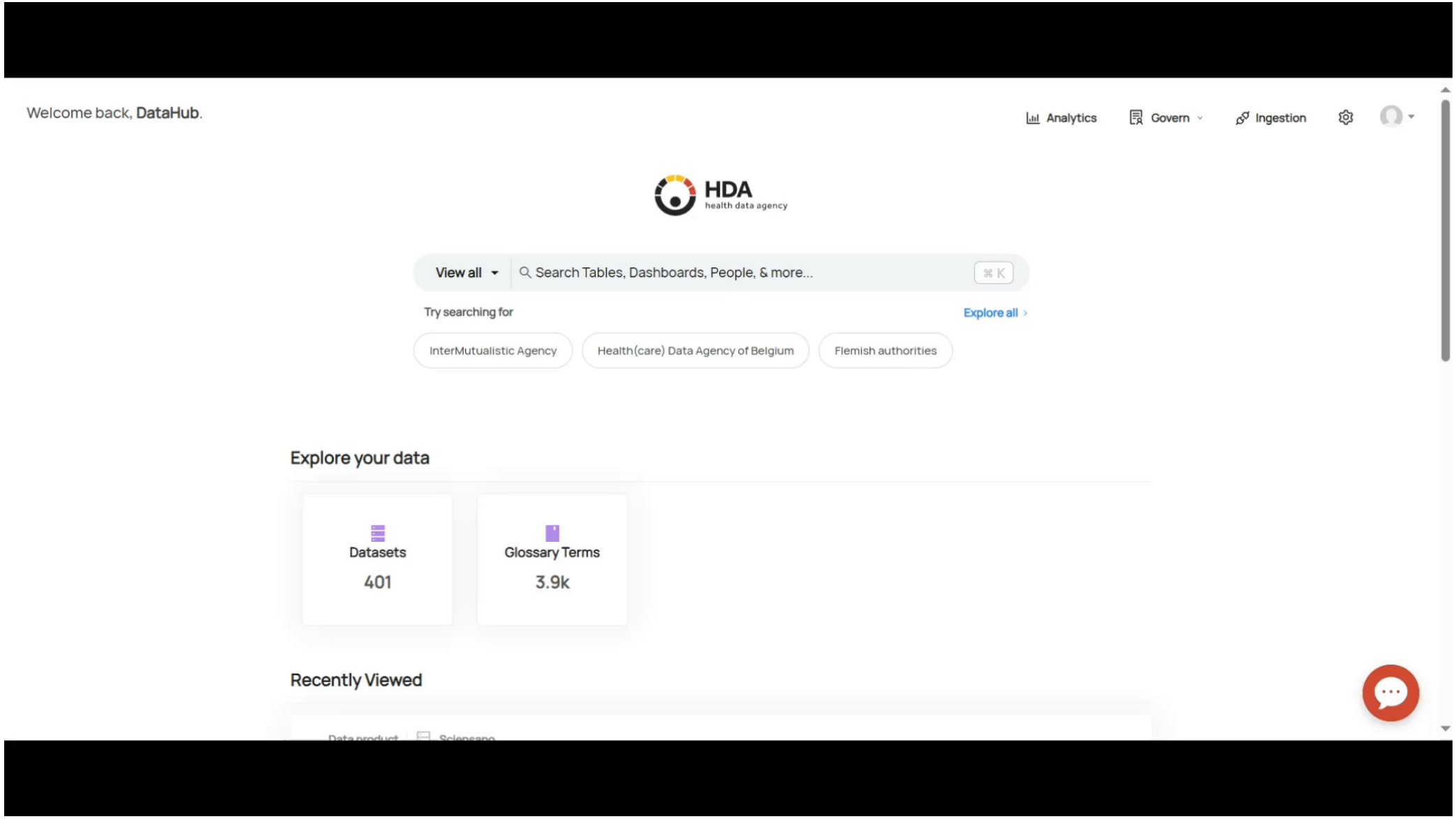
…Our goal is to continuously improve the way end-users can access and use (meta)data, in a faster and more efficient way.

Sneak peek of HDA's new AI catalogue assistant



HDA

Sneak peek of HDA’s new AI catalogue assistant



Conclusion

Towards a bright future

DataHub

Collibra

Q&A

We value your feedback!

Question 1: what feature(s) would you like to see added to or improved in the data catalogue?

Question 2: for which data domain should we collect and document more datasets in priority?


HDA