metaphacts

Boehringer Ingelheim

Knowledge Democratization

with an Enterprise Knowledge Graph at Boehringer Ingelheim

Interview with Maksim Kolchin, Knowledge Graph Platform Lead, Boehringer Ingelheim

Daniel Herzig:
Let‘s hear firsthand from Maksim Kolchin. Maksim is the Knowledge Graph Platform Lead at Boehringer Ingelheim, where he implemented a decision intelligence platform on top of over 10 knowledge graphs with us.

Maksim, can you tell us a little bit about the data management challenges that you were trying to address and that motivated you to build such a platform? What were the goals that you were trying to achieve at Boehringer Ingelheim?

Maksim Kolchin:
I have learned that pharma companies face similar issues when it comes to data management. As we know, the amount of data grows constantly; data is handled differently and within their own systems at each stage in the drug development pipeline; and data is not published or reused as well as it could be. So we could have a better data publishing and data reuse culture.

This results in a situation where data scientists, data analysts, and researchers have to spend their time on bringing data together for data analysis or for answering specific business questions. And like this, each use case basically generates new data without data reuse.

This results in a situation where data scientists, data analysts, and researchers have to spend their time on bringing data together for data analysis or for answering specific business questions. And like this, each use case basically generates new data without data reuse.

A couple of years ago, we started working on a new program at Boehringer Ingelheim to address these challenges. The program is called dataland and my team – the enterprise knowledge graph team – is part of this program. The goal of the program is to build a data-centric culture at Boehringer Ingelheim, address the challenges that I mentioned before, and implement the technical systems that are needed to tackle these challenges.

Daniel:
Maksim, you talked about topics like data centricity, bringing data together, reusing data. These are all aspects that I‘ve also heard when looking at previous approaches to data management. What is different in the knowledge graph based approach and why did you choose this approach compared to previous approaches like data warehousing?

Maksim:
Let‘s compare the data warehouse and data lake approaches with the knowledge graph approach on the conceptual level. The first difference that I would mention is the **data model**, specifically the way the data model is defined. In the knowledge graph approach, the model is defined together with the business and there is no separation between the conceptual, logical, and physical layers. So the model you define is just one thing: you define concepts, relationships, and attributes together with the business and you use the same concepts when you write queries; data analysts, data scientists, and then business analysts use these same concepts when they write queries. And this model is not somewhere outside of the data, but it lives together with the data and the data actually references this data model directly.

The second aspect refers to the **explicit links** that are defined between data points in a dataset. In a knowledge graph, we have global identifiers, which are, first of all, globally unique. Secondly, these are not just identifiers, they are not just numbers, but they also describe the protocol to access the data behind the identifier. So this results in explicit links.

The next important aspect is **federation** across different knowledge graphs and different data silos. This federation is possible thanks to SPARQL which is the query language and the HTTP-based protocol at the same time.

Additionally, it is not a secret that the graph data model is more **flexible** than the tabular model. You can extend your graph data model while you are improving your data, which means that you can ingest new data before you need to explicitly define the data model. You can massage and transform your data to the data model you have.

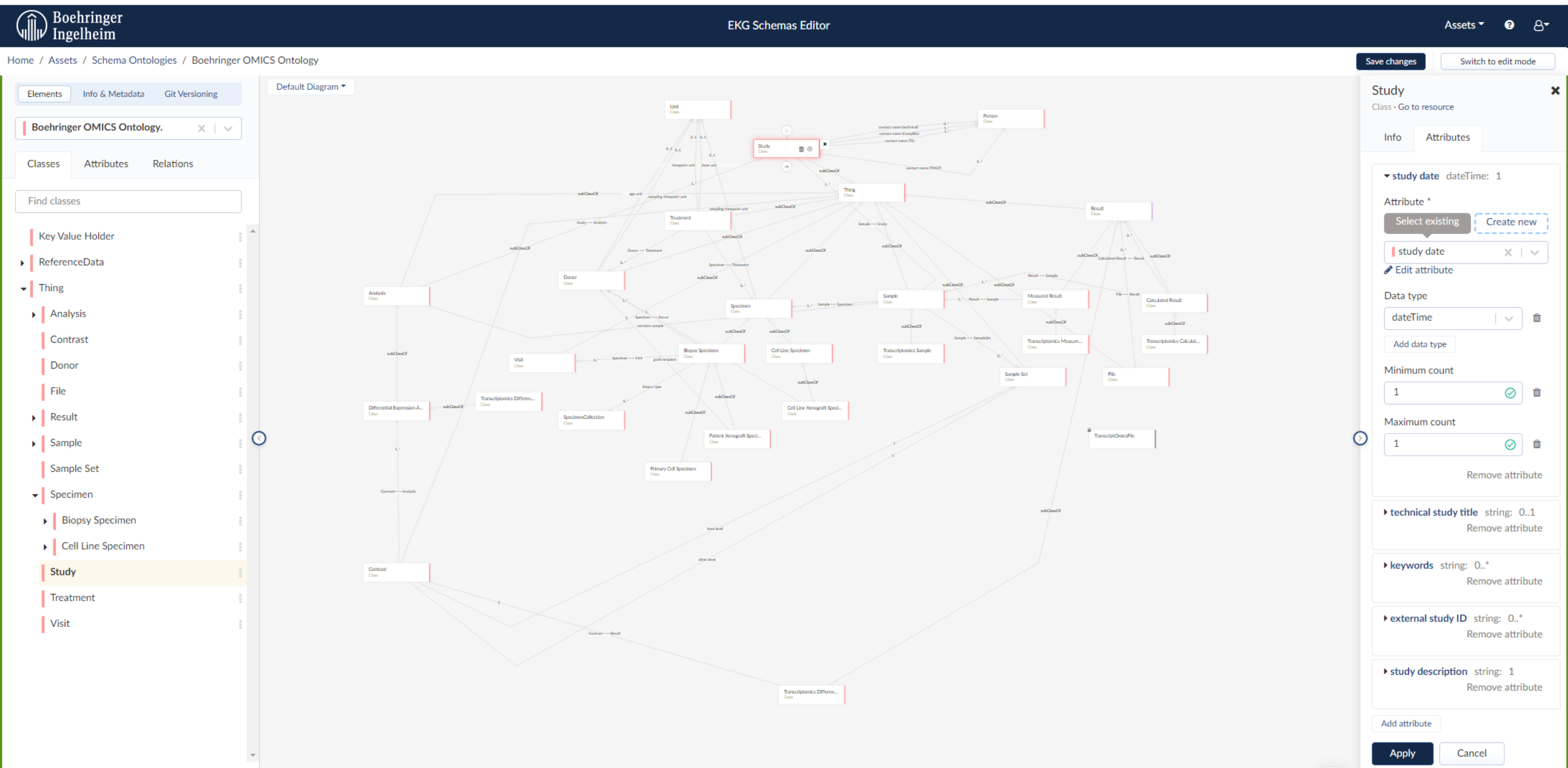
> *[The] graph data model is more flexible than the tabular model. You can extend your graph data model while you are improving your data, which means that you can ingest new data before you need to explicitly define the data model.*

The last point that I would mention is **data virtualization**. You don‘t need to always copy data to the knowledge graph. You can virtualize it. If you have data somewhere in Postgres, Oracle, or a Microsoft SQL server, you can describe the mapping between the tabular model and the graph model, and the engine will use this mapping at query time.

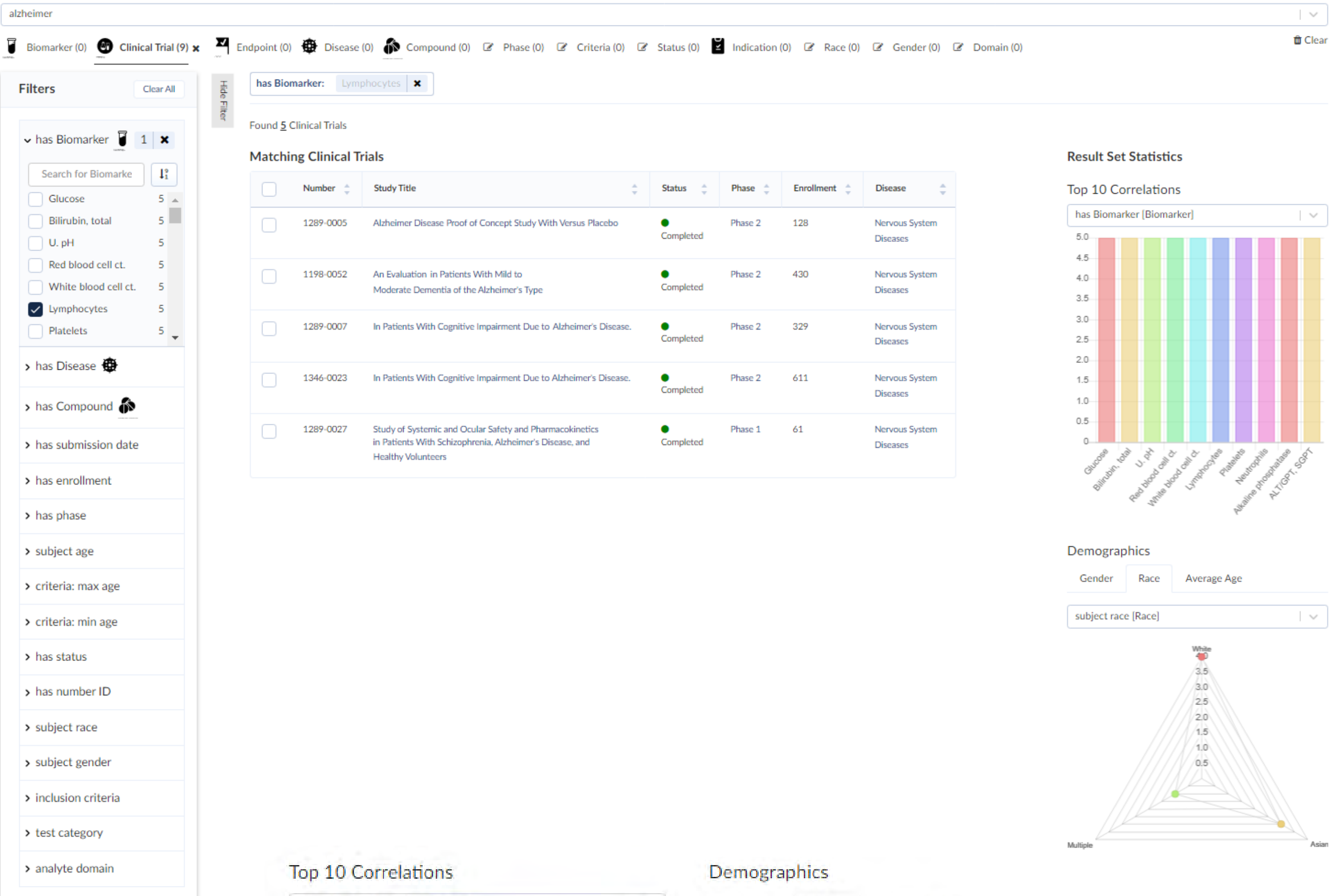
Daniel:
You mentioned quite a number of different aspects to the previous approaches. But if you now go one step back: What were the key elements that you considered when building such a platform? What were the key components that you took into account?

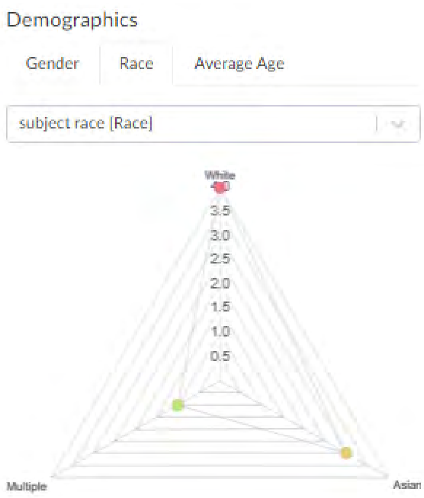
Maksim:
Working with knowledge graphs requires you to think about things which you didn‘t consider in the past, or in the data warehouse approach, for example. Specifically, it requires you to think about how you will deal with relationships between data points. You need to have a way to **create identifiers and reuse them across knowledge graphs.** For that, you probably need a smart tool, but you also need organizational approaches and processes in place to tackle this.

Another component which is required is a **collaborative tool for ontology and taxonomy management.** You need to bring business to this activity because IT folks lack the domain expertise, of course, and they cannot create this data model for you. So business needs to have a good tool for that. Here we leverage the metaphacts offering and we started using metaphactory’s visual modeling interface where users can create ontologies and taxonomies in a collaborative way. We just started and we are still learning, but it looks promising.



The last component that I would mention is a tool which allows data consumers with or without IT background to really **explore data across knowledge graphs.** We are talking about identifier reuse and linking data silos together, but we still have tools which provide views on top of one single database. So this component that I‘m talking about should allow people to look at the data across knowledge graphs, without boundaries. Users shouldn‘t need to reconnect from one database to another and the user experience should not have any boundaries between data sources, but should leverage the benefits of the explicit links between knowledge graphs. Having such a tool makes it easier to explain why you need explicit links. We need explicit links because they allows us to explore data.





Here we also use metaphactory because it has a special mode which allows you to federate across knowledge graphs. We built a use case agnostic application where users can leverage a keyword search to find an entity or data point they want to start their exploration with. When they find it, they get a 360 view over this entity – let's say a compound or a clinical trial. From here they can start exploring to other nodes until they answer their business question. With this data exploration tool it is easier to show the benefits of the knowledge graph approach.

Daniel:
In your answers so far, you several times mentioned multiple knowledge graphs and how they connect. In the past, I have talked about the enterprise knowledge graph. Now you mentioned that there are several knowledge graphs at Boehringer Ingelheim. Could you elaborate a little bit on this? What does it mean that you have several knowledge graphs and how do they connect?

Maksim:
Yes, as of today, we have more than **seven use cases** at different stages of development at Boehringer. We have a couple of them in production and others are in the ideation or POC stages. Together, these knowledge graphs form an enterprise knowledge graph.

From our perspective, an enterprise knowledge graph is a set of domain specific knowledge graphs and a set of shared ontologies and taxonomies developed by the business. All these things together form an enterprise knowledge graph. And, of course, the knowledge graphs need to share identifiers. Otherwise, they will be data silos, which we, of course, want to avoid.

One of the first use cases we addressed with the knowledge graph approach was around **publishing data coming from several laboratories** at Boehringer; specifically, publishing data in such a way that data scientists can actually bring these data together and perform analyses. In the past, the data resided in their own systems in each laboratory. Now, data scientists have a single place where they can get all data together. And it‘s not just another data silo, because the data is published following the FAIR principles and following the knowledge graph approach, which means that we can now can leverage the benefits we discussed before. Data scientists can take data coming from laboratories and easily mix it with data coming from other places.

In the past, the data resided in their own systems in each laboratory. Now, data scientists have a single place where they can get all data together. And it‘s not just another data silo, because the data is published following the FAIR principles and following the knowledge graph approach.

We also have other use cases, for example in one specific use case we have bunch of different **IT systems** with data hidden behind them. The goal here is to mix and provide context for this data, like data about tickets, incidents, IT systems, IT system leads, and so on.

Another use case deals with **document management** and the knowledge graph solution allows researchers to find existing documents that are relevant to the documents they are working on.

Daniel:
You mentioned several use cases that you have already realized at Boehringer. One of them was the use case about laboratory data – the use case which we started our collaboration with. If you think back at this first use case from which then the enterprise knowledge graph step by step evolved and grew, what were some lessons learned that you could share? Are there any recommendations that you can share with those who are just starting out with a knowledge graph approach?

Maksim:
There are two key learnings that I would like to share. Initially, the use case team created the data model – the ontologies and taxonomies – together with the consultants from metaphacts. That was the first iteration they worked on. But after some time, they created a second iteration and significantly improved the data model – and they did this themselves, without help from the consultants. So our learning here is that you probably don't always need an ontologist in the team. You probably do need such a role in the beginning but then the team can learn how to do data modeling, how to build ontologies, and how to define concepts and relationships between them. These were business folks that we are talking about – data stewards, data domain owners. They worked together on this and they succeeded. Now, they use this second iteration, this second version of the ontology.

> *[Our] learning here is that you probably don't always need an ontologist in the team. You probably do need such a role in the beginning but then the team can learn how to do data modeling, how to build ontologies, and how to define concepts and relationships between them.*

As you may have guessed, at Boehringer we use metaphactory and the second learning was that building custom applications with metaphactory for our use cases was relatively easy without too much effort or resources required.

Our goal as a support team is to provide tools and services for business use cases, for data providers, so that they can focus on bringing data together, improving the quality of this data, cleaning it, and so on. And this should be their main job. The rest should be done for them. The rest are these services which we provide and one of our learnings was that offerings for use case agnostic applications are important. So our team has built use case agnostic applications so that the data providers can focus on data creation and data management instead of building and maintaining custom applications.

Daniel:
Okay, Maksim, we have heard that you have established this new platform and this new tooling to get a knowledge graph to work at Boehringer, but were there any cultural or organizational changes that were required as well? Have you somehow had to adapt to this new technology also on the organizational level at Boehringer?

Maksim:
That's a good question because technology cannot solve all issues and the knowledge graph approach also requires organizational changes. Actually, not only the knowledge graph approach. When you try to improve data management in the organization, you usually need a cultural change as well.

> *[Technology] cannot solve all issues and the knowledge graph approach also requires organizational changes. [...] When you try to improve data management in the organization, you usually need a cultural change as well.*

In the dataland program, apart from building technological capabilities, we also work on defining new roles and responsibilities. I would mention two roles here. The first one is **data domain owners** – the people who actually need to own data and need to be responsible for the data within their own domain. They are accountable for this data and need to think about how to publish it better and how to make the data better for the data consumers.

> *In the dataland program, apart from building technological capabilities, we also work on defining new roles and responsibilities: data domain owners and data stewards.*

The second role is **data stewards** – the people who actually work together with IT and data engineers on finding data sources, bringing these data together, and aligning them to a domain model. They are responsible for the data and they are part of each use case team. The challenge here is actually implementing these roles in the organization and creating processes that work. This is what we are trying to achieve in dataland.

Daniel:
If you now look into the future, what are your plans? What are the next steps that you're going to take? And where do you think your knowledge graph journey will take you from here?

Maksim:
As a company we have already spent quite some time on this initiative but I would not yet say that we've achieved everything we wanted and we need. We still have a long journey ahead. I see a lot of potential in the technology and in the approach, but it's not just about technology, but also about changes in the organizational culture and structure, which take time. We have seen quick benefits but I believe we will see even more benefits in the long term. We could talk about decision intelligence and machine learning when looking towards the future, but even the easiest things like data exploration or the possibility to search across data silos are already great achievements in my opinion.

> *I see a lot of potential in the technology and in the approach, but it's not just about technology, but also about changes in the organizational culture and structure, which take time. We have seen quick benefits but I believe we will see even more benefits in the long term.*

Daniel Herzig-Sommer, COO, metaphacts

Maksim Kolchin, Knowledge Graph Platform Lead, Boehringer Ingelheim

GET IN TOUCH

metaphacts GmbH
info@metaphacts.com
www.metaphacts.com

@metaphacts

metaphacts-gmbh

metaphacts

Watch the full interview:
» **https://www.youtube.com/watch?v=NswT7m_iqj4**

Demo
Open Knowledge Graphs
» **https://wikidata.metaphacts.com**

Get started
Get started with metaphactory for free and start building your Knowledge Graph today!
» **https://metaphacts.com/get-started**