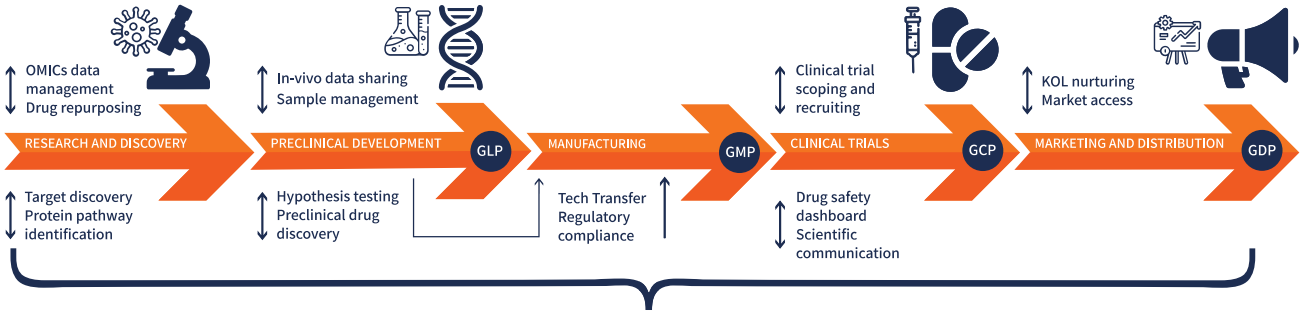
Authors: Ilian Uzunov, Todor Primov

In the last decade, we've seen a massive surge of knowledge graph technology. This is especially true for Life Sciences, Pharma and Healthcare, which are central to Ontotext's portfolio. Why is that? The first and most important thing is that knowledge graphs offer a completely new perspective on how you can manage and consume your data. There is a huge difference between a company that operates with data locked in various data stores and one that easily navigates through its data, driven by its business needs.

By interlinking concepts and placing your proprietary knowledge in the context of other relevant knowledge, graphs enable fast data consumption and allow you to navigate through them in multiple ways, depending on your use case. For example, if you are in Drug Safety, you can track the different relationships between a treatment and a specific drug or you check reported adverse events. If you are in Drug R&D, you may want to find what specific proteins are targeted by a specific drug or check molecular interactions data, gene expression data, etc.

Knowledge Graph Use Cases Across the Pharma Value Chain

Pharma is one of the most active adopters of knowledge graph technology in the industry sectors we cover and we have considerable experience working with some of the most knowledge intensive enterprises in this domain.



We have worked in ***research and development*** with use cases like therapeutic target

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

discovery, drug repurposing or OMICS data management. In ***preclinical development*** we have customers using our technology to build knowledge graph powered applications for hypothesis testing, preclinical drug discovery, etc. In drug ***manufacturing*** our technology powers use cases in regulatory compliance, technology transfer, etc. We have built strong expertise in ***clinical trials*** with use cases in drug safety, scientific communication and regulator intelligence. Finally, when it comes to ***marketing and distribution***, we have worked on use cases for key opinion leader nurturing, medical inquiries, analytics as well as product labeling and updates.

Our Methodology for Building Knowledge Graphs

Ontotext's methodology for building knowledge graphs is based on 20+ years of delivering multiple projects in various domains.

Let's go briefly over the steps!

1. Compiling A List of Competency Questions

The first step determines the scope of your project and how it will develop. Starting by compiling a list of competency questions puts you in the shoes of different stakeholders and allows you to view the same questions from various angles.

For example, a competency question like: "I would like to know which are the key opinion leaders in a specific research medical domain" will reflect the point of view of a Clinical Researcher, which might be completely different from the needs of a Clinical Trial Project Manager.

2. Identifying Relevant Datasets

After you've defined the competency questions you want to address with your knowledge graph, the next step is to select the datasets for it. To be able to determine which datasets are most relevant to you, you need to have a good understanding of the domain and a clear idea of what you want to achieve.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

3. Analyzing Selected Datasets

Once you have your candidate datasets, you can start analyzing them more thoroughly. What kind of content will they bring in? What is the number of records available? How comprehensive is a particular dataset? What are the available identifiers you can use to interlink it with other data sources? Which dataset will provide which piece of data to solve which specific question?

Going deeper and deeper, you can identify the schema, disambiguate and further define the meaning of concepts and get a better understanding of the knowledge encrypted in the datasets.

4. Data Cleaning and Normalization

While most datasets in Life Sciences, Pharma and Healthcare are high quality datasets that don't require much cleaning and normalization, there are still things that have to be done.

For example, there are mismatched identifiers, URIs that refer to wrong namespaces, etc. You need to fix all these errors and discrepancies so that when you load your data in the knowledge graph, it can be automatically resolved.

5. Data Enrichment

In the next step, we integrate our Natural Language Processing pipelines, populated with some of the referential datasets, ontologies, terminologies, vocabularies, etc., which are key in the specific domain.

We use these resources to identify objects mentioned in the text in order to disambiguate terms, avoid having information locked in textual fields, etc. As to complexity, it can range from simple Name Entity Recognition to Relation Extraction from small snippets coming from some of the metadata-rich datasets.

6. Data Conversion and Model Harmonization

Now that you have a great variety of data, you need to harmonize it. You also have to

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

transform some of the data into native formats like XML, JSON, TSV into RDF, which means that you need to have data conversion scripts in place.

7. Deduplication and Instance Matching

Another common problem is that, often, the data you want to use is in overlapping datasets. As there are identical entities in different sources, you have to deduplicate and match them semantically.

For example, both NCBI Gene and Ensembl provide information about gene and genomics data. So, you need to semantically match the different instances across the two datasets, provide a unified representation of this information and limit the redundancy of the data.

8. Creating Data Update Procedures

Once you have your knowledge graph, if you want to keep it live, you have to create an automated data update procedure. This includes steps 4 to 7 and ensures not only that these steps are automated, but that the required artifacts are generated.

Another thing you need to introduce is data quality checks and validations, as your schema will evolve and there will be other changes. So, you need to ensure that each new release of a dataset is consistent with your data model.

9. Query Services: Answers to Competency Questions

Now that all is set, you can query your knowledge graph in multiple ways, starting with a basic SPARQL query. If the previous steps have been implemented successfully, your SPARQL query should be able to answer the competency questions you had defined in the beginning of the project.

10. Serve Data to Analytical Dashboards

Finally, when you have all this in place, you can develop a complex semantic search and other deep analytics tools on top of your knowledge graph through dashboards or other algorithms.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Authors: Ilian Uzunov, Todor Primov

Ontotext's LinkedLifeData Inventory: an Overview

Based on our solid methodology and extensive use case experience in the domain, we have gone one step further to facilitate our clients. We have created a LinkedLiveData Inventory of 200+ public datasets and ontologies in Life Sciences, Pharma and Healthcare.

As you can see in the diagram below, the public datasets in this Inventory span across various domains like genomics, proteomics, metabolomics, pathways and interactions, clinical, medical and various types of scientific publications, etc.

Another important part of the Inventory are the referential ontologies and terminologies that we use. Some of them are already distributed in an RDF format (like, for example, SNOMED CT), but many others (like MedDRA, ICD-9, ICD-10, etc.) are not. We keep all ontologies in our Inventory in RDF and use them to semantically normalize the data for all other datasets like PathwayCommons, FDA, National Drug Code product reports (NDC), Drugs@FDA and many others.

Copyright: Ontotext New York | Sofia | Basel

Ontotext's LinkedLifeData Inventory: Use Cases

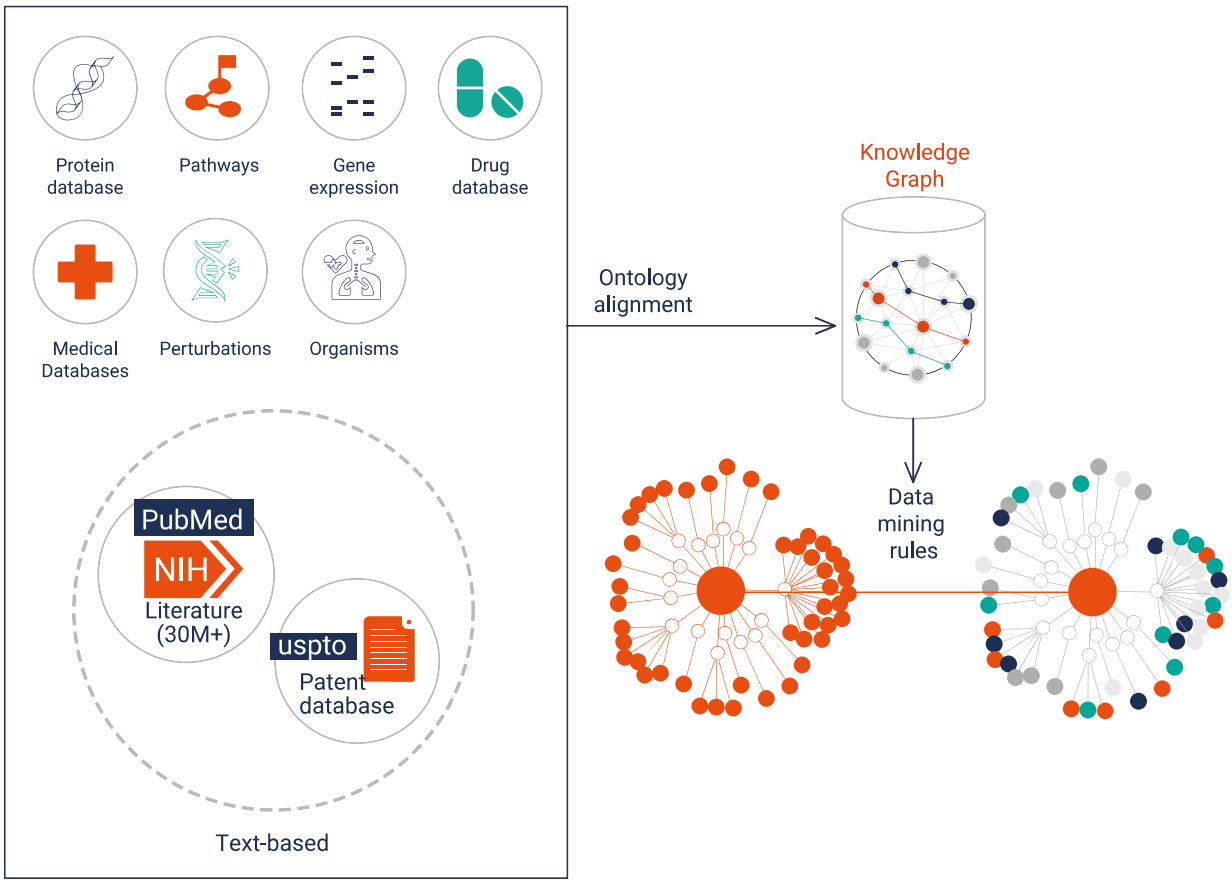
The LinkedLifeData Inventory was created to help you in your Life Sciences, Pharma and Healthcare projects and there are different ways to get it. You can subscribe to a custom set of datasets and preferred update frequency (DaaS). You can let us develop, manage and maintain your custom knowledge graph (KGaaS). Or you can rely on us to develop your custom knowledge graph powered solution for semantic search and analytics based on your specific needs.

Let's have a look at two examples where Ontotext's LinkedLifeData Inventory helped a client achieve a particular goal.

A Top 5 Global Pharma Company Preclinical Discovery Knowledge Graph

Our first example shows how the LinkedLifeData Inventory was recently used in a project for a top 5 Global Pharma company. The company wanted to build a multimodal preclinical knowledge graph that would provide highly interlinked information across various domains. The datasets we used were: UniProt(Human), Ensembl (Human), Reactome, StringDB, ChEMBL (+ core ontologies), GWAS and Expression data.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com



The result was a knowledge graph of more than 1,300,000,000 facts (with 38% inferred new facts), linking all different categories of information and providing insights on top of it. This graph enables users to define complex queries scattered across different datasets as well as to get insights from the dashboards and data exploration UI functionalities, which we have built together with our partners at metaphacts.

NuMedii New Drug Discovery in Linked Data

Our second example shows how our LinkedLifeData Inventory was used in a project for NuMedii. NuMedii wanted to build an intelligent analytical solution supporting research activities related to identification of new therapies for treating idiopathic pulmonary fibrosis (IPF).

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

We applied the same methodology and a much larger dataset collection of more than 20 public datasets (including UMLS, PubMed, MeSH, UniProt, PubChem, DrugBank, FAERS, NCBI, etc.) as well as NuMedii's proprietary data. We also augmented and enriched the knowledge graph with information that was not present in the structured data sources, but came from unstructured texts in scientific journals.

The result was an expert knowledge graph of 7.98 billion triples, which powered NuMedii analytical solution, covering genomics, proteomics, metabolomics, disease conditions, drug

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

products, scientific literature and various biomedical ontologies. This graph enables users to identify patterns and correlations between biomedical concepts and test new hypotheses facilitated by a dynamically enriched knowledge graph.

Wrap Up

In both of the above examples, using Ontotext's LinkedLifeData Inventory and our proven methodology made it possible to significantly shorten the time for delivery compared to starting the implementation from scratch.

Thanks to the automated data update process, each dataset is automatically updated, semantically normalized, enriched and validated towards its previously defined schema. The result is data that is easily findable, accessible, interoperable and that can be reused in other projects (FAIR).

So, whatever you need for your specific use case, we have it in our LinkedLifeData Inventory and we can use it to enrich your proprietary data. The efficient combination of this Inventory and our technology allows us to deliver purpose-built business solutions and do it in no time.

Want to take advantage of Ontotext's LinkedLifeData Inventory?

Book a Discovery Call

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com