Common Infrastructure for National Cohorts in Europe, Canada, and Africa

- CINECA -

Deliverable D1.2 - Query expansion service

Work Package:	WP1 - Federated Data Discovery and Querying
Lead Beneficiary:	European Molecular Biology Laboratory
WP Leaders:	Jonathan Dursi (SickKids/UHN) and Jordi Rambla de Argila (CRG)
Contributing Partner(s):	SickKids, CRG, HES-SO, UCT
Contractual Delivery Date:	31st December, 2020
Actual Delivery Date:	18 December, 2020
Authors of this Deliverable:	Romain Tanzer (HES-SO) Nona Naderi (HES-SO) Douglas Teodoro (HES-SO) Anais Mottaz (HES-SO) Patrick Ruch (HES-SO) Jonathan Dursi (SickKids) Jordi Rambla de Argila (CRG)
Reviewed by:	Melanie Courtot (EMBL-EBI)
Approved by:	Thomas Keane (EMBL-EBI)
Dissemination Level:	Public
Type of Deliverable:	Demonstrator
Grant agreement:	No. 825775 Horizon 2020 (H2020-SC1-BHC-2018-2020)
Type of action:	RIA
Start Date:	1 Jan 2019
Duration:	48 months

Table of contents:

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

1. Executive Summary

CINECA aims to support federated queries and analyses of distributed cohorts across continents. But human health datasets are extremely diverse; many different types of data are collected for many different kinds of health studies by many different health research communities. As a result, different cohort datasets often use different ontologies to describe similar kinds of entities, or represent concepts, such as genomic variation differently.

CINECA must span this diversity of data representations in order to achieve its goals of connecting health research cohort data. The work of WP3 partially addresses discoverability of datasets by defining a standard minimal cohort-level data representation which will be common across all cohorts; but that does not address cohort-level data that falls outside of the minimal common data model, nor does it address the representation of patient-level data. WP1's role is to design and deploy API access to both cohort- and patient-level data, and a fundamental functionality of the infrastructure is to allow the user to find the appropriate dataset independently of the ontology used to map locally the different cohorts or indifferently of the format and syntax used to describe the variants.

Deliverable D1.2, Query expansion service, also supports the WP5 use cases and queries (from data set search up to variant interpretation), by implementing and demonstrating a query expansion service API that improves findability and searchability of distributed cohort data. Multiple kinds of query expansions are available for enabling further data integration and interoperability, including horizontal expansion, i.e., across ontological systems, and vertical expansion, i.e., within sublevels of the same ontological resource.

This report describes the work done on query expansion.

2. Project objectives

This deliverable contributes to the following WP1 objectives:

a) To improve searchability and interoperability of cohorts using WP1 APIs
b) To enable the user to use free text to search over WP1 APIs
c) To provide a web service to support CINECA upstream use cases, particularly WP5.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

3. Detailed report on the deliverable

3.1 Background

For researchers to be able to make federated queries or analysis across multiple cohorts and data set end points, there is a need for a mechanism to translate their queries into the terminological and ontological resources used by each site to structure their data (Figure 1). WP1 supports an ontology agnostic approach: any user information request can be expanded. The use cases that we explore correspond to different situations: 1. the expansion is done using unique identifiers from ontological resources available in the Ontology Lookup Service (OLS); 2. the expansion is performed at the term level.

To ensure a common metadata ontology, WP3 has defined the Genomics Cohorts Knowledge Ontology (GECKO) [2] which defines the minimal metadata model to represent cohorts genomic and phenotypic data. Using GECKO for cross-cohort discovery requires each of the cohort data dictionaries be mapped to GECKO. In some cases, in particular where data is already encoded with a specific national terminological systems (e.g. ICD-10 GM, ICD-9 CM, ICD-11), the horizontal query expansion across equivalent ontology terms will greatly ease the process and alleviate the need for further mapping.

To ensure an exhaustive search and compatibility with existing ontologies, the query expansion plays an important role [3]. As shown in Figure 1, this service enriches user searches using ontologies, and a data driven approach, where search terms are expanded to similar concepts in other ontologies, and users are enabled to navigate through the ontology hierarchy.

The solution to share genotypic and phenotypic data adopted by WP1 is based on the Beacon specification. A reference Beacon v2.x implementation [4], developed by CRG, is publicly available for the different sites to use in order to share their data.

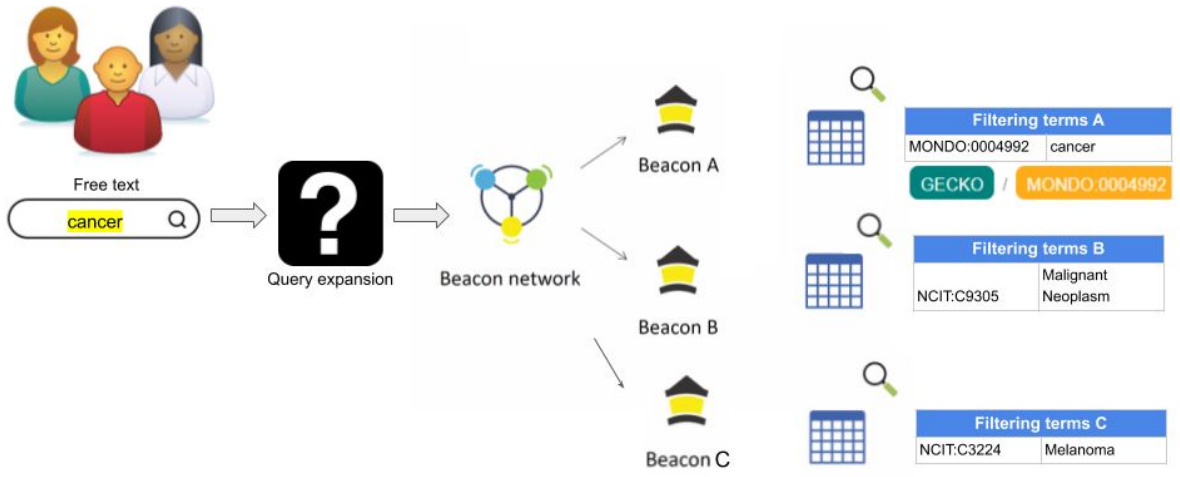


Figure 1 - *Example of the output of the query expansion API for the query: "cancer"*

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

As it is hard for the CINECA network to control how the data will be structured by each site, in order to increase compatibility, the query expansion service proposes multiple ontology terms to use in addition to the original query. The distributed query engine (Beacon network) can then use the public mapping description of the Beacon(s) to decide which ontological system to use for each site. Thus, it will increase the recall of datasets returned by the Beacon network. To increase search recall, the original queries can be translated, e.g., from free text to ontological concepts, and expanded, i.e., into subclasses or synonyms.

Another important aspect to consider is the level of granularity used to structure the cohort dataset versus the search keyword or ontology term used to query Beacon. For example, a researcher looking for cohorts with East Asian ethnicity demographics would need to query East Asian as well as all the subcategories from the Hancestro terminology [5] or some other ontology (see Figure 2). This kind of search without any help would be highly time consuming. In order to solve this issue, the query expansion should provide a way to expand terminologies vertically, that is, to subclass levels.

Figure 2 - *Hancestro vertical expansion*

Variant description is another important subject to deal with, as there are many ways to describe the same variant (see Figure 3). First, the user needs to select at which level the genetic variation is expressed: at the genome, protein, or transcript level. Each of these levels correspond to different forms to characterize a given variation. A genomic variant is described with regards to its position on the chromosome and may affect multiple genes/transcripts. A transcript variant (caused by alternative splicing) is described according to its position in the gene. And a proteomic variant may result from multiple different mutations on the gene or on the genome (redundancy or degeneracy of the genetic code). The combinatorial nature of these levels (many-to-many relationships) hinder a linear mapping between them. Second, there are multiple syntactic variations at each level of description. The query expansion service should be able to accept any variant query format and

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

convert it into a Beacon specific query (Figure 3). This will simplify the variant query for the user and facilitate the integration with external services by accepting their variant description.

http://goldorak.hesge.ch/synvar/

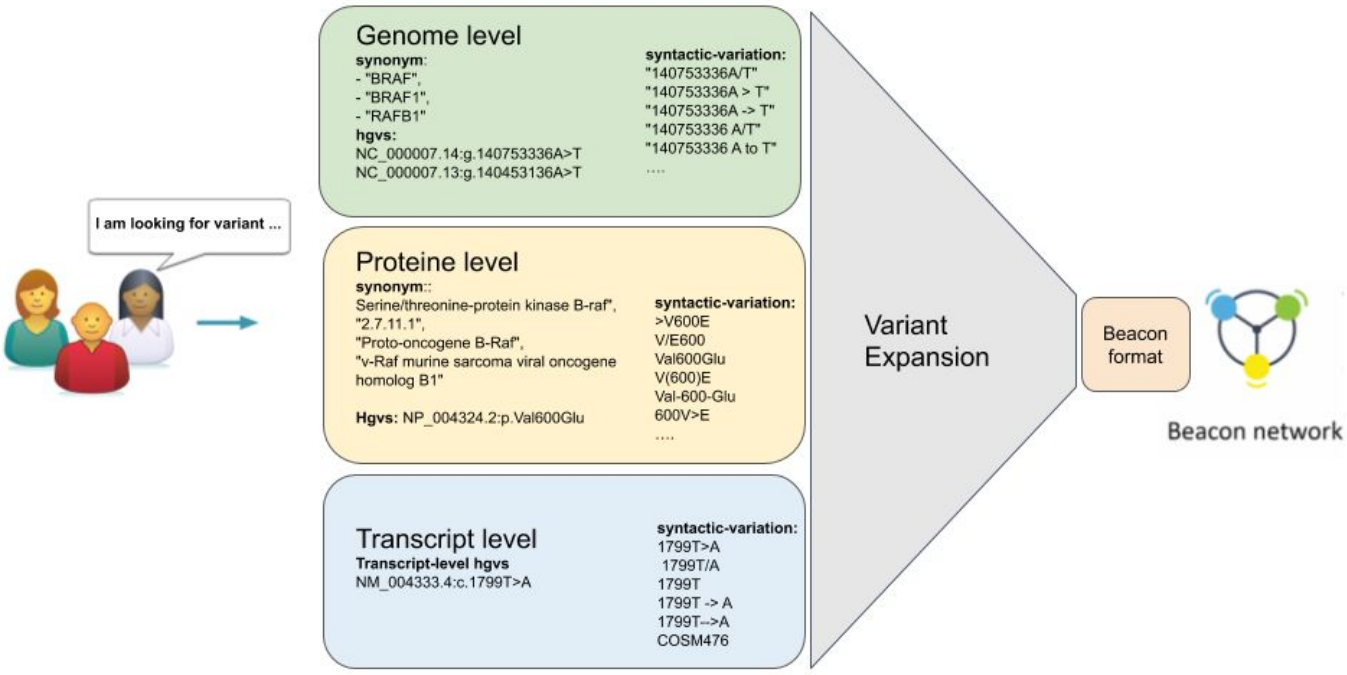


Figure 3 - *Variant expansion: independently of the user syntax, the query expansion module will identify equivalent synonyms that can be used to fetch local data in the Beacon(s).*

3.2 Work Done

For this demonstration, CINECA partners at HES-SO have implemented the query expansion service to be integrated with upstream work package use cases, particularly WP5. This demonstration shows an overview of the query expansion process and how it can be integrated in CINECA in the context of WP1.

Figure 4 provides a summary of all the expansion services provided by this deliverable. For each type of expansion, we have created and provided a different API endpoint.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

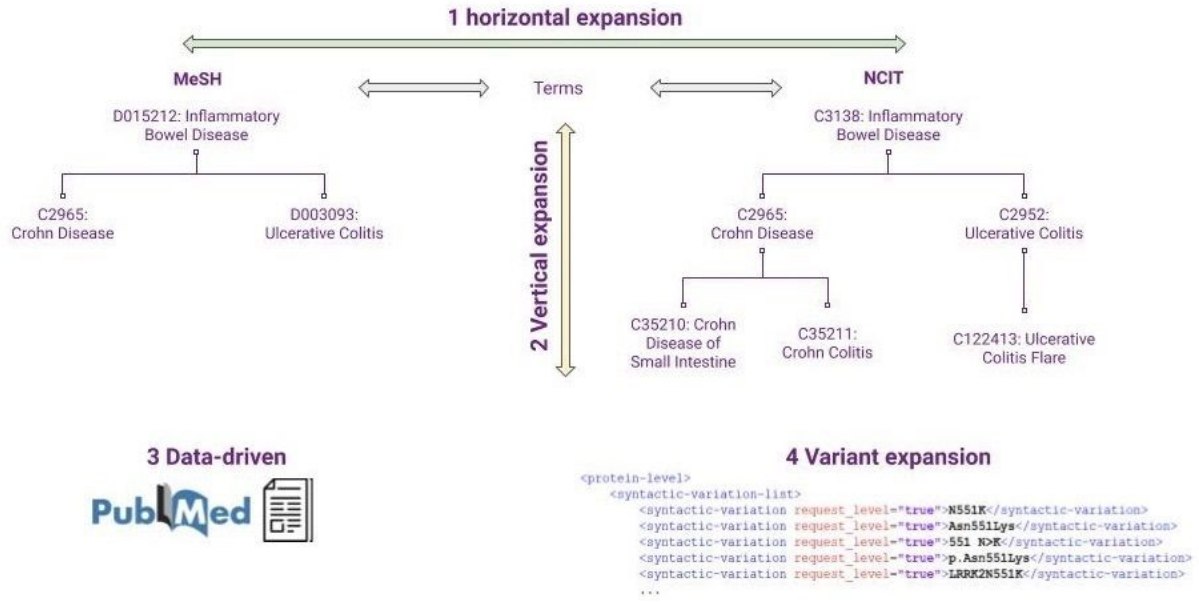


Figure 4 - *Query expansion overview*

Five API endpoints have been developed for this demonstration:

1. **Data driven expansion**

This endpoint provides a data driven expansion, which could serve as a suggestion of similar terms to users during their search. This expansion model relies on the word2vec algorithm [6,7] and a Medline collection of one million abstracts to train the generative algorithm. Our approach includes a query expansion module that computes the word embedding for query terms and performs their expansion using the k-nearest embedded word vectors [8]. As an example, “tumor” can be expanded to “cancer”. Unlike ontology-driven expansion, the model is likely to expand a concept to non-formal relationships. Thus an ontology would expand a concept with synonyms, generic or specific, while such a data-driven expansion is likely to also expand a given concept with antonyms or any related concepts in the broad sense: thus, *hypertension* could be expanded by *blood pressure* or *mmHg*. Hence, it will usually require end user-intervention to select the most relevant terms.

2. **Search ontology: OLS and UMLS**

As an initial prototype, this endpoint encapsulates the EBI Ontology Lookup service (OLS) search API, to return concepts in ontologies matching the text query. We also added an Unified Medical Language System (UMLS) exact match expansion, grouping the terminology by CUI Concept (CUI) [9,10].

3. **Horizontal expansion (cross reference)**

This endpoint provides a list of cross references using the EBI OLS API. It lists all the equivalent concepts from other terminologies loaded in OLS.

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

4. Vertical expansion OLS (hierarchy)

This endpoint provides a way to expand a terminology term through its relatives (either its parents or children) with the level parameter indicating how many levels the expansion needs to go in the specific ontology hierarchy.

5. Variant expansion

This endpoint relies on SIB-hosted SVIP\SynVar services [11] to generate an extensive variant query. Different variant forms are accepted as input: genome, protein, transcript, as well syntactic variation for these different levels of description.

These multiple expansions can be combined to support the discovery platform and the different upstream WPs. For instance, to query different Beacon endpoints (via Beacon Network) and find the relevant patients which had cancer, as in Figure 5, the combination of expansion presented in Figure 6 can be used. It describes how the initial search term ("cancer") can be expanded and then the expanded query can be sent to the Beacon network according to the different ontologies for the local Beacon endpoints. The Beacon network will use the exposed ontological endpoint description to decide which ontology to use.

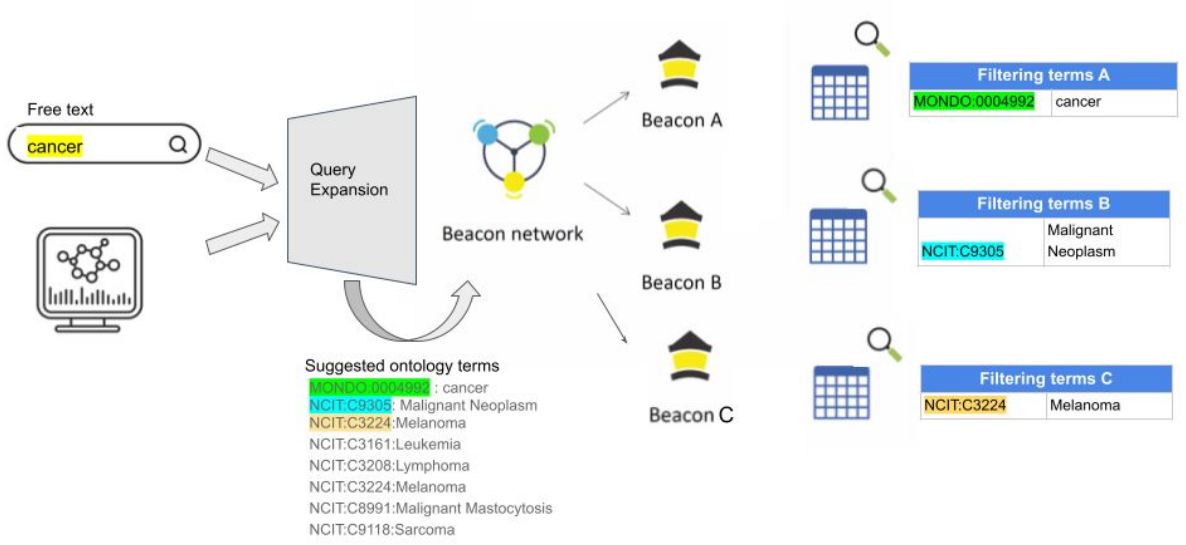


Figure 5 - *Query expansion suggestion of ontology terms*

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

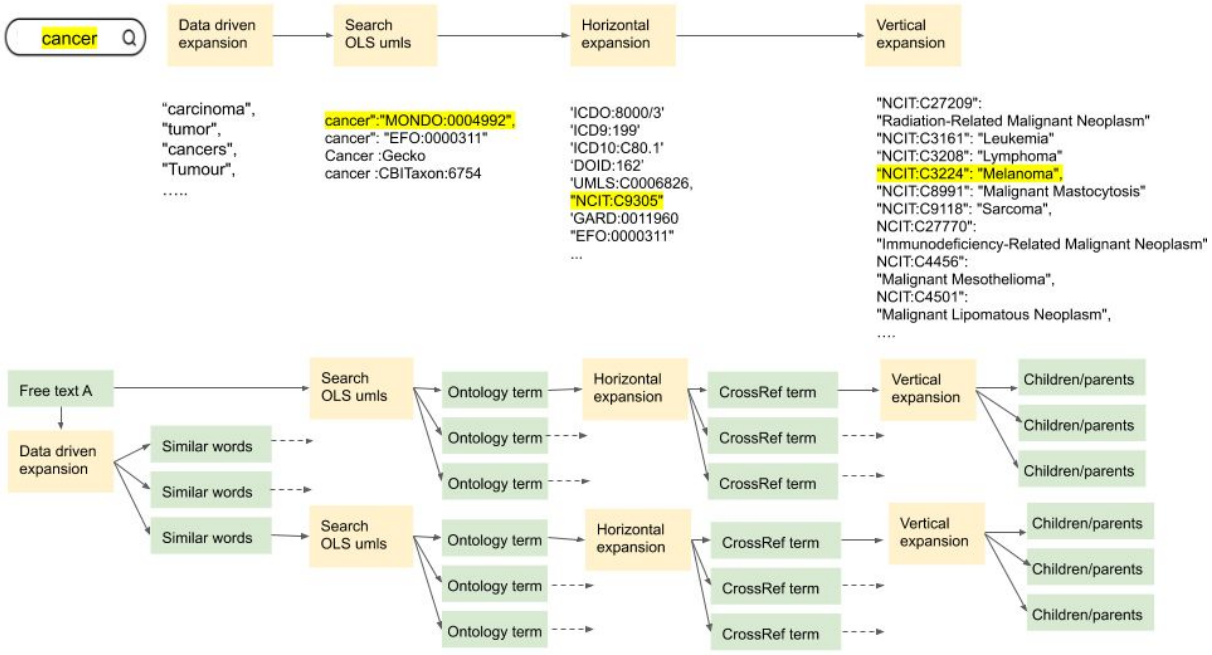


Figure 6 - *Example of query expansion: the highlighted terms indicate the sequence from an initial query ("cancer"), keyword based, followed by the different expansion modes described above (from MONDO to NCIT ontologies: "horizontal mode" and from NCIT:neoplasm (C9305) to NCIT:melanoma (C3224): "vertical mode"), until the actual ontological representation in the local Beacon.*

3.3 Demonstration

The demonstration consists of two pre-recorded webinars. First, the ontology query expansion video describes why we need query expansion and the role it plays in the larger efforts of WP1. The Query Expansion Ontology Expansion video is found here (Appendix 7.1). Second, the variant expansion video describes the variant expansion using SynVar[de] for integration with a Beacon genomic query. The Query Expansion Variant Expansion video is found here (Appendix 7.2).

The technical demonstration consists of six parts:

1. Query expansion context
2. Example of query expansion flow for a Beacon query
3. Data driven expansion API demonstration
4. Search ontology API demonstration
5. Horizontal expansion API demonstration
6. Vertical expansion API demonstration

The variant expansion use-case demonstration consists of three parts:

1. Context of variant expansion
2. Different input forms
3. Variant expansion API (converging to beacon format) demonstration

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

4. Conclusion and next steps

The CINECA query expansion service has been designed to be modular. It explores a wide range of expansion models: ontology-driven expansion (e.g. synonyms, generic, specific), as well as expansion based on broader association models, which build on neural text mining approaches.

The services are exploiting the flexibility of the Beacon v2 protocol and syntax. Different expansions can be combined in different orders to match the needs of individual CINECA use cases including variant expansion in WP5 to more generic conceptual expansion for virtually any dataset search task.

In future work, we plan to assess the efficiency of the different expansion models across a variety of queries of increasing complexity. Finally, we would like to quantify the improvement of the query expansions regarding retrieval effectiveness (recall and precision). Therefore, we are considering evaluating the impact of these services on some ad hoc data set search tasks, as explored by the BioCaddie evaluation campaign and benchmarks [8, 12].

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

5. List of abbreviations

Abbreviation	Definition
API	Application Programming Interface
GA4GH	The Global Alliance for Genomics and Health (https://www.ga4gh.org), a standards body for health genomics APIs and data models.
UI	User Interface
WP1	Work package 1 of the CINECA project, implementing federated data discovery and analysis queries
UMLS	Unified Medical Language System (https://www.nlm.nih.gov/research/umls/index.html), is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.
OLS	Ontology Lookup Service (https://www.ebi.ac.uk/ols/index), a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions.
MeSH	Medical Subject Headings (https://www.nlm.nih.gov/mesh/meshhome.html) thesaurus is a controlled and hierarchically-organised vocabulary produced by the National Library of Medicine used for indexing, cataloging, and searching of biomedical and health-related information.
GECKO	Genomics Cohorts Knowledge Ontology, an ontology to represent genomics cohort attributes. To browse: (https://www.ebi.ac.uk/ols/ontologies/gecko), and for a description of GECKO's development: (https://github.com/IHCC-cohorts/GECKO)
CRG	Centre for Genomic Regulation

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

6. References

1. Beacon API specification v1.0 (2019). Available at: https://github.com/ga4gh-beacon/specification. (Accessed: 17h May 2020).

2. Genomics Cohorts Knowledge Ontology < Ontology Lookup Service < EMBL-EBI. https://www.ebi.ac.uk/ols/ontologies/gecko. (Updated: Dec. 2020).

3. Bhogal J., Macfarlane A., Smith P. (2007). A review of ontology based query expansion. Information Processing & Management, 43(4), 866-886. DOI: 10.1016/j.ipm.2006.09.003.

4. EGA Archive Beacon v2.x (2020). Available at: https://github.com/EGA-archive/beacon-2.x/ (Accessed: 8 June 2020).

5. Genomics Cohorts Knowledge Ontology < Ontology Lookup Service < EMBL-EBI. https://www.ebi.ac.uk/ols/ontologies/hancestro. (Updated: Nov. 2020).

6. Bengio Y., Ducharme R., Vincent P. et al. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3, 1137-1155. Part of the Studies in Fuzziness and Soft Computing series, 194. DOI: 10.1007/3-540-33486-6_6.

7. Mikolov T., Chen K., Corrado G. et al. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv preprint, arXiv: 1301.3781.

8. Teodoro D., Mottin L., Gobeill J., et al. (2017). Improving average ranking precision in user searches for biomedical research datasets. Database (Oxford), 2017, bax083. DOI: 10.1093/database/bax083.

9. Côté R., Reisinger F., Martens L., et al. (2010) The Ontology Lookup Service: bigger and better. Nucleic Acids Research, 38(Web Server issue), 155-160. DOI: 10.1093/nar/gkq331.

10. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32(Database issue), 267-270. DOI: 10.1093/nar/gkh061.

11. Caucheteur D., Gobeill J., Mottaz A. et al. (2020). Text-Mining Services of the Swiss Variant Interpretation Platform for Oncology. Studies in Health Technologies and Informatics, 270, 884-888. DOI: 10.3233/SHTI200288.

12. Roberts K., Gururaj A., Chen X. et al. (2017) Information Retrieval for Biomedical Datasets: The 2016 bioCADDIE Dataset Retrieval Challenge. Database (Oxford), 2017, bax068. DOI: 10.1093/database/bax068.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

7. Work Packages in CINECA

WP1 - Federated Data Discovery and Querying
WP2 - Interoperable Authentication and Authorisation Infrastructure
WP3 - Cohort Level Meta Data Representation
WP4 - Federated Joint Cohort Analysis
WP5 - Healthcare Interoperability and Clinical Applications
WP6 - Outreach, training and dissemination
WP7 - Ethical and legal governance framework for transnational data-sharing
WP8 - Project Management and coordination
WP9 - Ethics requirements

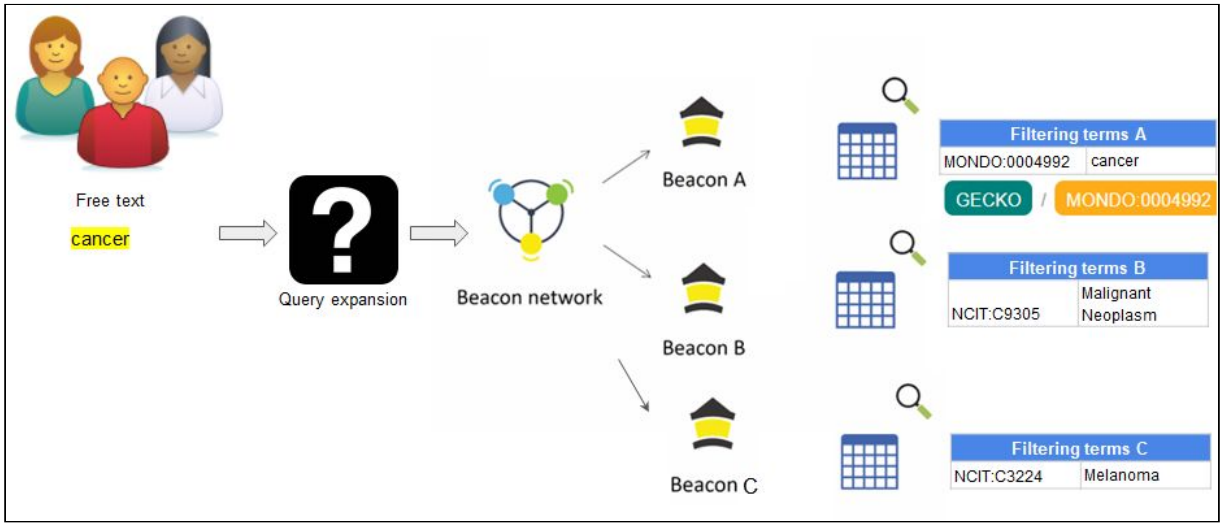
8. Delivery and schedule

The delivery is on time.

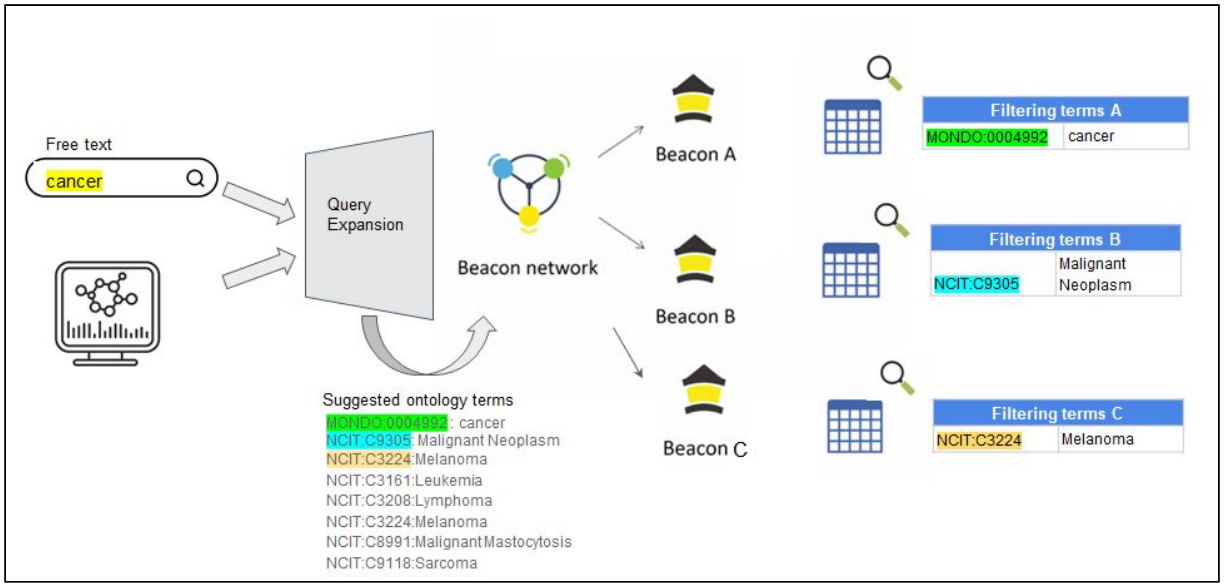
9. Appendices

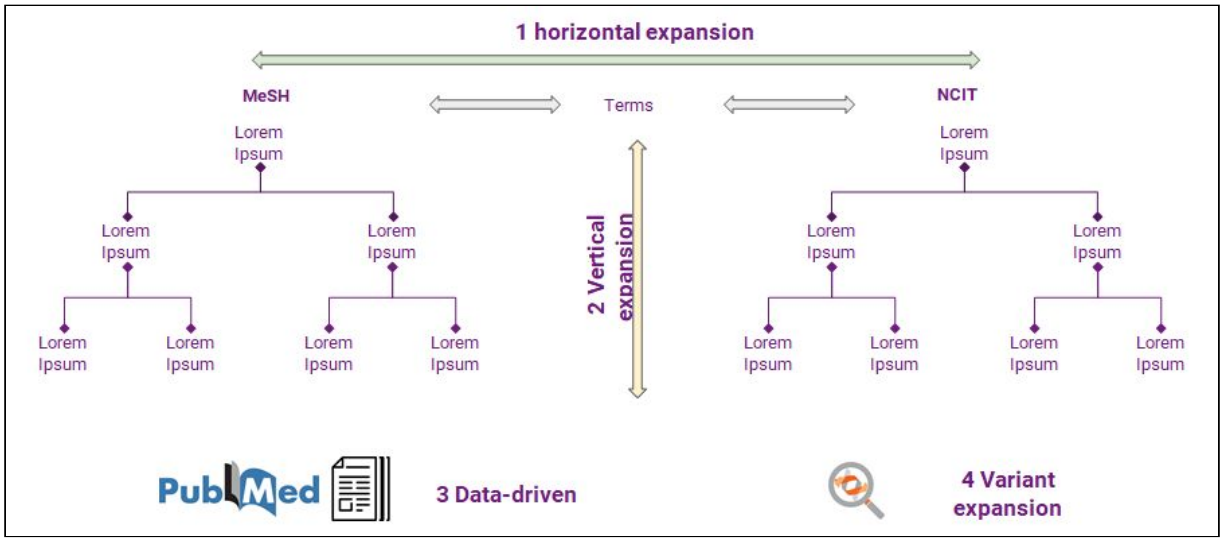
9.1 Appendix 1 - A Catalogue For CINECA Services, Introduction

This appendix includes a summary of the slides presented in the introductory video for the Query expansion service.

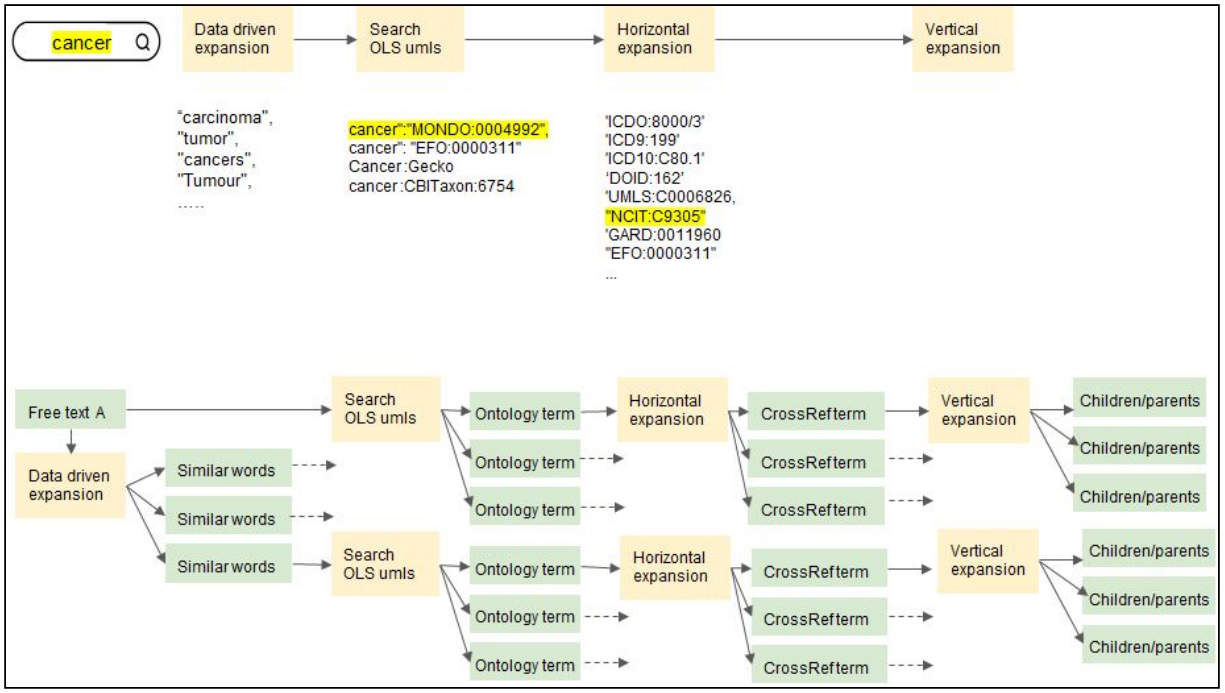


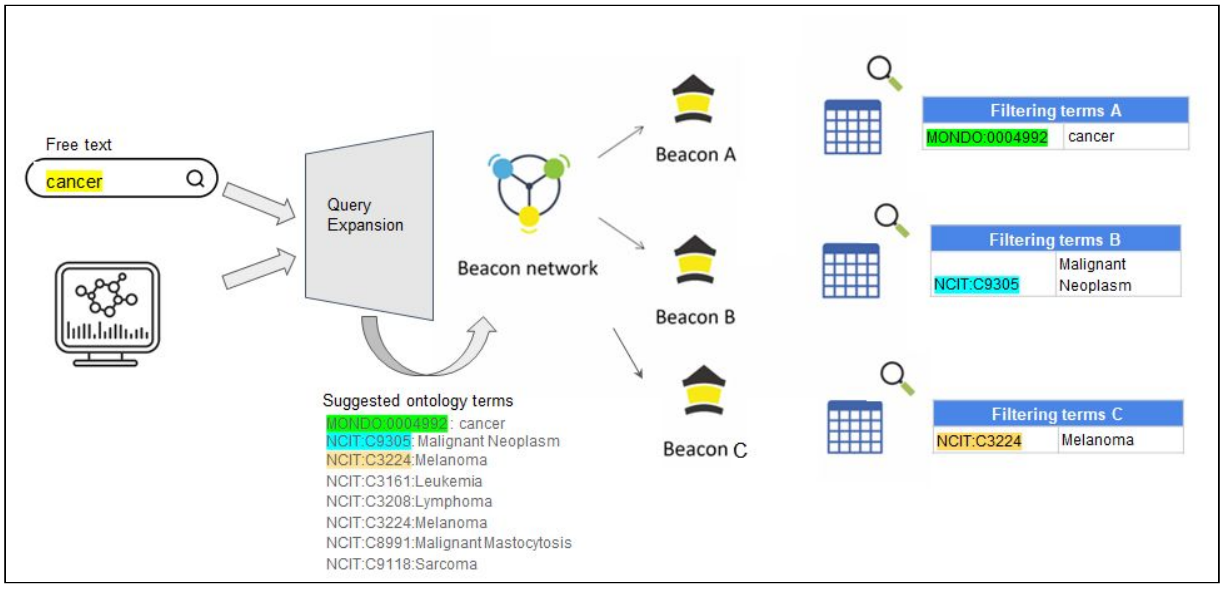
This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.





This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

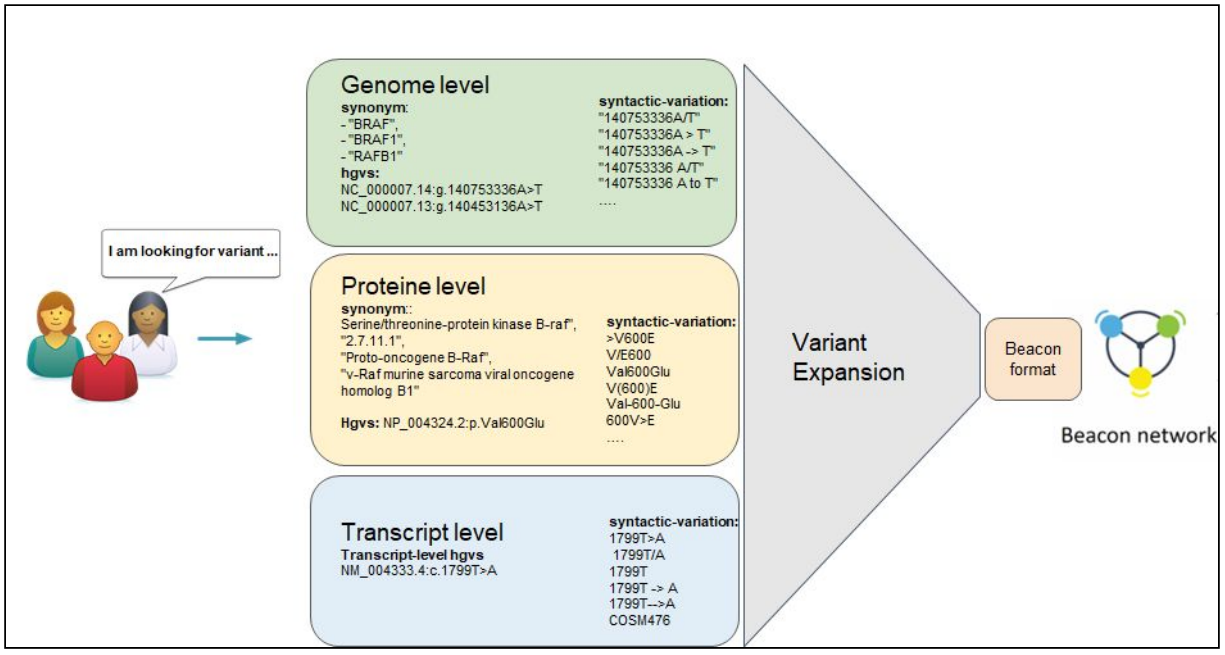


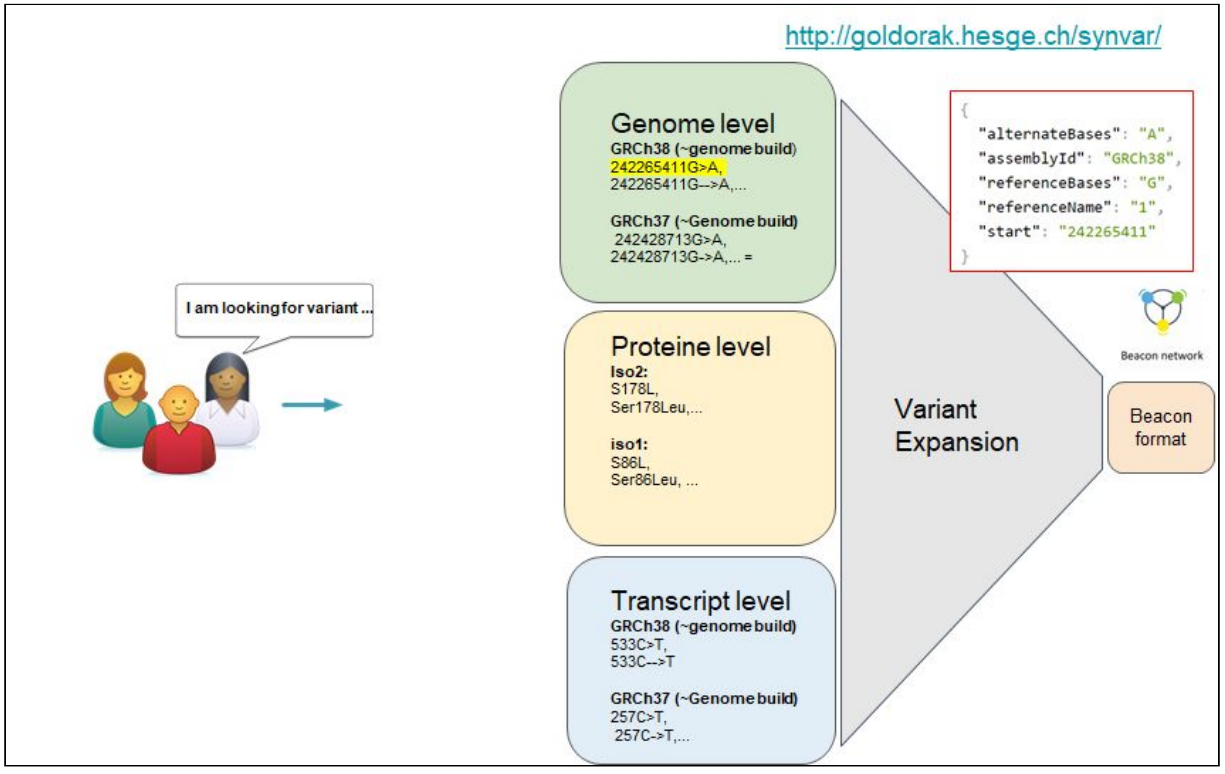


This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

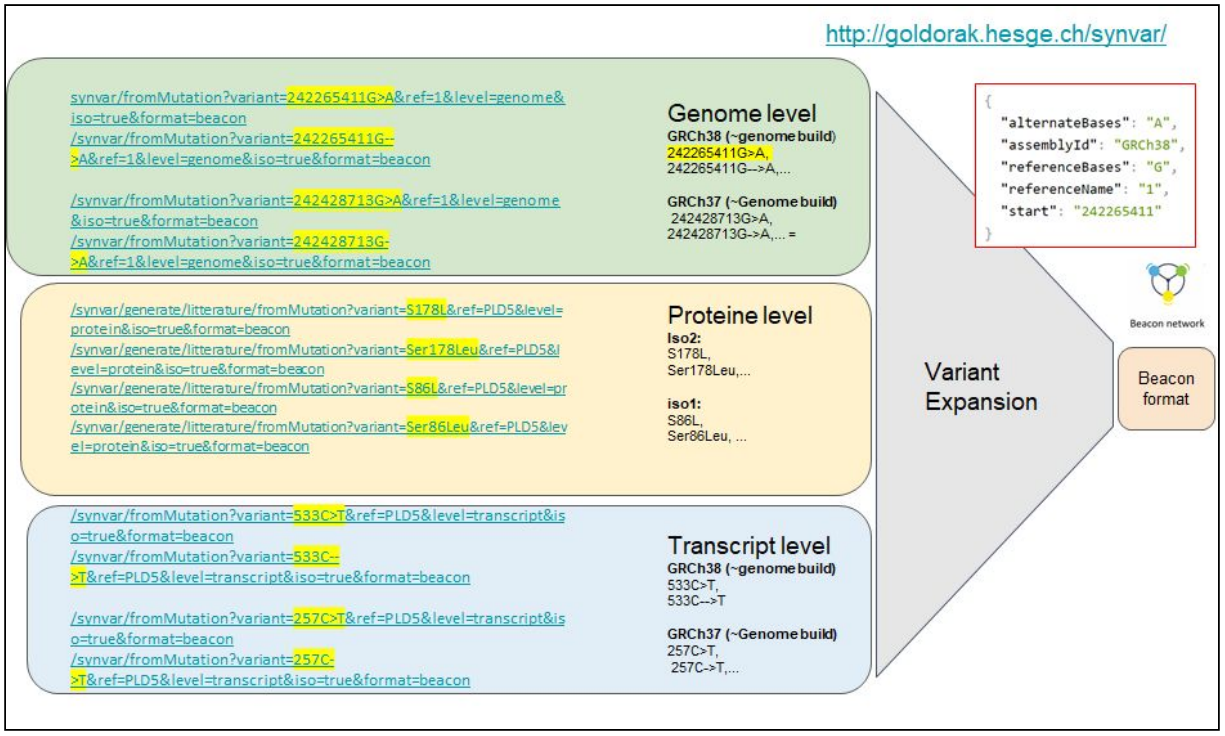
9.2 Appendix 2 - A Catalogue For CINECA Services, Variant expansion perspective

This appendix includes a summary of the slides presented in the technical walkthrough demo video for the Query expansion service, from the variant-specific perspective.





This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.



This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.