

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

D3.2 - Semantic and harmonisation best practice

Work Package:	WP3 - Cohort Level metadata Representation
Lead Beneficiary:	European Molecular Biology Laboratory
WP Leaders:	Fiona Brinkman (SFU), Melanie Courtot (EMBL-EBI)
Contributing Partner(s):	UMCG, SFU, UCT, HES-SO
Contractual Delivery Date:	30th June, 2021
Actual Delivery Date:	21st June, 2021
Authors of this Deliverable:	Melanie Courtot (EMBL-EBI), Isuru Liyanage (EMBL-EBI)
Contributors:	Justin Cook (SFU), Tony Burdett (EMBL-EBI), Leslie Glass (EMBL-EBI)
Reviewed by:	Claudia Vasallo Vega, Mamana Mbiyavanga, Patrick Ruch
Approved by:	Thomas Keane (EMBL-EBI)
Dissemination Level:	Public
Type of Deliverable:	Other
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:

1. Executive Summary	3
2. Project objectives	3
3. Detailed report on the deliverable	4
3.1 Background	4
3.2 Best practices for cohort metadata harmonisation	5
3.2.1 Selecting a minimum metadata model for cohort data	5
3.2.2 Automated mapping of cohort dictionaries to semantic model	5
3.2.3 Registration of harmonised cohort metadata in a cohort registry	6
3.3 Semantic modeling	7
3.3.1 GECKO development	7
3.3.2 GECKO content	9
3.3.3 GECKO hosting	10
3.4 Results	12
3.4.1 Federated cohort discovery in CINECA	12
Mapping CINECA synthetic cohort data to GECKO	12
Federated discovery of cohort data	12
3.4.2 Federated cohort discovery beyond CINECA	13
The IHCC cohort atlas	13
The Davos Alzheimer's Collaborative Atlas	14
3.5 Next steps	14
3.5.1 Interoperability with Maelstrom	15
3.5.2 Engagement with other communities	15
4. References	16
5. Abbreviations	16
6. Work Packages in CINECA	16
7. Delivery and schedule	16
8. Appendices	17
8.1 The IHCC Cohort Atlas: faceted browsing and discovery using GECKO	17
8.2 Resources imported in GECKO	24



1. Executive Summary

To support human cohort genomic and other omic data discovery and analysis across jurisdictions, basic data such as cohort participants' demographic data, diseases, medication etc. (termed “minimal metadata”) needs to be harmonised. Individual cohorts are constrained by size, ancestral origins, and geographic boundaries that limit the subgroups, exposures, outcomes, and interactions which can be examined. Combining data across large cohorts to address questions none of them can answer alone enhances the value of each and leverages the enormous investments already made in them to address pressing questions in global health. By capturing genomic, epidemiological, clinical and environmental data from genetically and environmentally diverse populations, including populations that are traditionally under-represented, we will be able to capture novel factors associated with health and disease that are applicable to both individuals and communities globally.

We provide best practices for cohort metadata harmonisation, using the semantic platform we deployed in the cloud to enable cohort owners to map their data and harmonise against the GECKO (GENomics Cohorts Knowledge Ontology) we developed. GECKO is derived from the CINECA minimal metadata model of the basic set of attributes that should be recorded with all cohorts and is critical to aid initial querying across jurisdictions for suitable dataset discovery. We describe how this minimal metadata model was formalised using modern semantic standards, making it interoperable with external efforts and machine readable. Furthermore, we present how those practices were successfully used at scale, both within CINECA for data discovery in WP1 and in the synthetic datasets constructed by WP3, and outside of CINECA such as in the International HundredK+ Cohorts Consortium (IHCC) and the Davos Alzheimer's Collaborative (DAC). Finally, we highlight ongoing work for alignment with other efforts in the community and future opportunities.

2. Project objectives

WP3 Task 3.2 objectives:

1. To define current semantic coding standards.
2. To develop semantic mapping and data harmonisation strategies in support of analysis WPs.
3. To deliver a semantic toolkit supporting semantic mapping(s).



3. Detailed report on the deliverable

3.1 Background

Progress in sequencing technologies and omics data generation has facilitated the appearance of human genomics cohorts of ever increasing size. Those cohorts can generate new insights into the causal relationship between a genotype and observed phenotype, provided relevant datasets can be identified and integrated to support further analysis (Lee et al. 2019). There are many challenges regarding cohort diversity, as well as standards for their discovery and representation. Several international projects are aiming to tackle those. For example, All of Us¹ in North America specifically targets diverse populations, encompassing a wide range of ethnicities and socio-economic statuses. Many national genomics programs are emerging such as the Genome India² project launched in July 2019, which aims to catalogue the genomic data of over 10,000 Indians. The diversity of cohort projects as well as the heterogeneous sources generating those datasets mean that while data becomes available at scale, its discovery, interpretation and further re-analysis remains a challenge.

CINECA's goal is to enable federated queries and analyses of the varying and wide-ranging datasets from the [CINECA cohorts](#) as shown on Figure 1. To achieve this, each of the cohorts must:

1. Select a minimal metadata model that supports interoperability with other cohorts
2. Map cohort-specific dictionaries to the selected minimal metadata model
3. Register their cohort in a central cohort registry, such that they can be discovered and searched from a central cohort browser

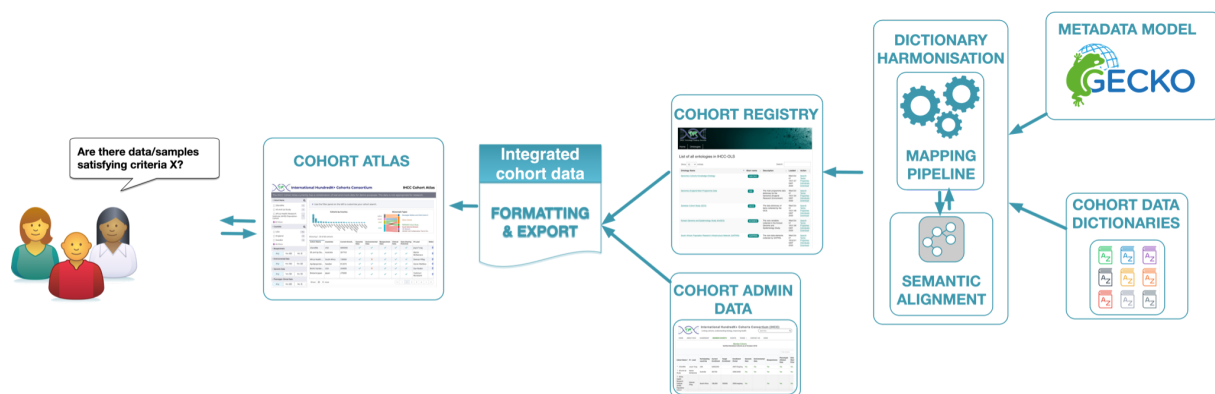


Figure 1. Overall pipeline enabling federating discovery of human cohort of interest.

¹ <https://allofus.nih.gov>

²

<https://economictimes.indiatimes.com/news/science/india-to-launch-its-1st-human-genome-cataloguing-project/articleshow/70323116.cms>

This deliverable defines best practices for CINECA cohorts to select a minimal metadata model, converting it into a machine readable semantic standard, and deploying a platform for adoption of the standard across cohorts. **Together, the semantic model and corresponding infrastructure provide a complete solution for cohort metadata harmonisation.** Cohort harmonisation enables federated discovery of datasets of interest through deployment of cohort atlases such as in the IHCC and the DAC described in [section 3.4.2](#).

3.2 Best practices for cohort metadata harmonisation

This section provides best practices and examples for cohort owners to harmonise their cohort metadata.

3.2.1 Selecting a minimum metadata model for cohort data

A semantic representation based on the minimal metadata model is needed to standardise the diverse variables from each cohort's dataset, and thus be harmonised to enable federated querying. While some work was done in [D3.1 Cohort minimal metadata model](#)³ to identify attributes shared between cohorts, there was no semantic representation or best practice to deploy those shared attributes in CINECA and beyond. In CINECA, the current cohorts did not use any semantic framework, and we therefore developed the GECKO (GEnomics Cohorts Knowledge Ontology) application ontology, a shared semantic model to which cohorts are mapped. GECKO has been used to ensure the interoperability of cohort variables within CINECA and beyond. [Section 3.5.2](#) discusses further opportunities to align and reuse different models.

3.2.2 Automated mapping of cohort dictionaries to semantic model

We have improved our mechanisms to add more cohort dictionaries, by providing spreadsheet templates for cohort owners to represent their metadata attributes in, which can then be automatically converted into the standard OWL representation and loaded into the registry. Those mechanisms leverage some of our automated mapping tools such as Zooma⁴ for text-mining or OxO⁵ for cross-references (See section [3.3.3](#)), to suggest mappings to the cohort owners. Suggestions are annotated with further metadata such as confidence levels and description to aid in decision making. Figure 2 below depicts the template spreadsheet populated with suggestions alongside other supplementary data. Suggestions with high confidence are automatically selected as the mapped attribute, whereas other suggestions are listed for manual curation. While we do not anticipate being able to fully support automated mapping to the CINECA model, we believe based on our experience in other projects that there are opportunities to increase the impact of manual curations. This could be done by adding them to our ZOOMA knowledgebase, thereby making them available for

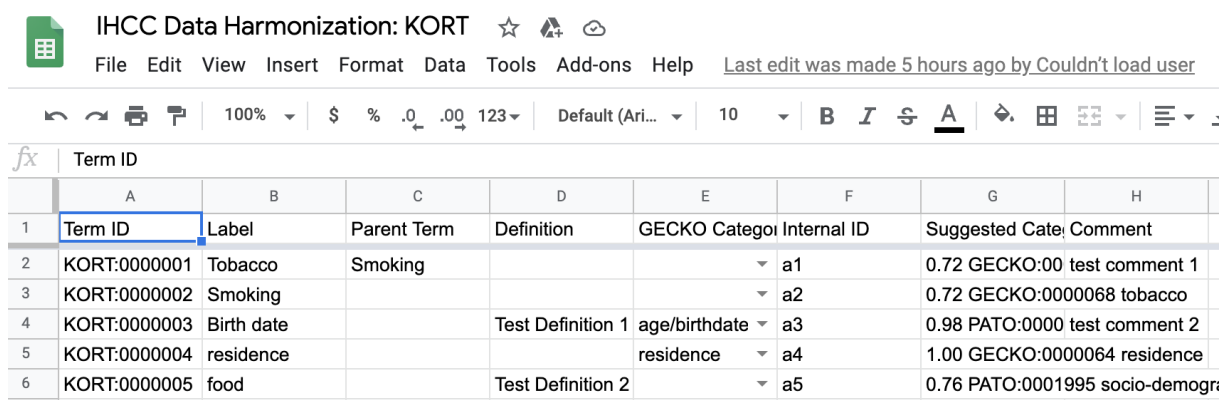
³ <https://doi.org/10.5281/zenodo.4575460>

⁴ <https://www.ebi.ac.uk/spot/zooma/>

⁵ <https://www.ebi.ac.uk/spot/oxo/>



the next submissions. Additionally this workflow is clearly defined and implemented in software to streamline the cohort integration process.



	A	B	C	D	E	F	G	H
1	Term ID	Label	Parent Term	Definition	GECKO Category	Internal ID	Suggested Category	Comment
2	KORT:0000001	Tobacco	Smoking			a1	0.72 GECKO:00	test comment 1
3	KORT:0000002	Smoking				a2	0.72 GECKO:0000068 tobacco	
4	KORT:0000003	Birth date		Test Definition 1	age/birthdate	a3	0.98 PATO:0000	test comment 2
5	KORT:0000004	residence			residence	a4	1.00 GECKO:0000064 residence	
6	KORT:0000005	food		Test Definition 2		a5	0.76 PATO:0001995 socio-demogr	

Figure 2. Template populated with suggestions from automatic mapping tools.

3.2.3 Registration of harmonised cohort metadata in a cohort registry

To streamline access to the converted data dictionaries, as well as facilitate addition of new cohorts, we have set up an instance of the Ontology Look-up Service (OLS)⁶. The OLS is a repository for biomedical ontologies that provides a single point of access to the latest ontology versions, as well as a simple web interface and a rich set of restful APIs to access its data. We have repurposed OLS as a repository of cohort data dictionaries where all the cohort dictionaries are stored. This provides human readability and search of harmonised cohort dictionaries as shown in Figure 3, as well as interoperability with our semantic toolkit for annotation and mapping as described in [section 3.3.3](#). It also includes powerful REST APIs that can be used by a cohort browser to retrieve the harmonised cohort data.

⁶ <https://www.ebi.ac.uk/ols/>

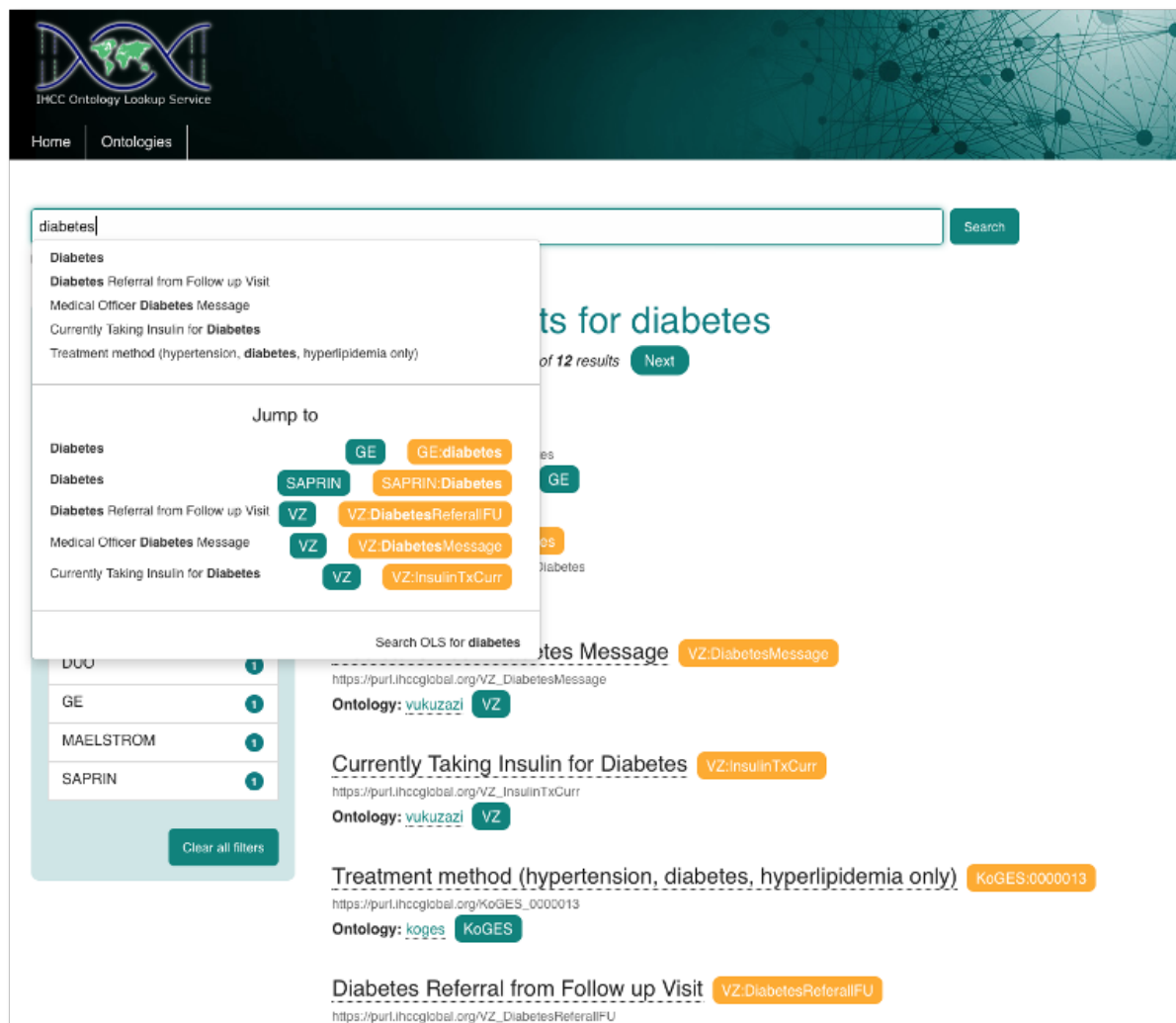


Figure 3. The cohort registry which centralises harmonised cohort dictionaries.

3.3 Semantic modeling

This section provides more details about the GECKO model development and content.

3.3.1 GECKO development

We have leveraged the work done in CINECA to define a [minimal metadata model](#). The model was distributed as a spreadsheet, which was useful for human review but not compliant with latest semantic standards such as stable identifiers or versioning, nor amenable to machine readability. To address this we used the model as the basis to create the GECKO application ontology for consistent representation of cohort metadata. The minimal metadata model had taken into account existing semantic efforts: some cohorts and tools in CINECA and other EuCan projects were already using the high-level categories from

Maelstrom⁷. Those high-level categories were further expanded to reflect the deeper granularity of CINECA cohorts. When developing GECKO, we attempted to reuse the Maelstrom categories natively but were faced with two different issues: (1) Maelstrom categories are distributed as YAML without unique identifiers (2) the CC-BY-ND licence⁸ the categories were released under prevented building a derivative, i.e., reusing and extending the model. Consequently, we have created GECKO classes that map to Maelstrom categories at high level, with the aim of replacing them with native Maelstrom classes as soon as possible. We have had fruitful discussions with the Maelstrom team and this is pending imminent resolution as they have agreed to update their licencing terms. In addition to efforts on reusing Maelstrom categories, care was taken to ensure GECKO follows the Open Biological and Biomedical Ontology (OBO) Foundry⁹ principles for development. This helps interoperability with additional resources such as described in 3.2.3. GECKO standardises the CINECA minimal metadata model by creating an OWL file with identifiers, versioning, licensing etc. GECKO also enriches the CINECA model by importing classes from other reference ontologies, such as CMO¹⁰ or OBI¹¹. Those imported classes further specify the GECKO hierarchy, and offer additional opportunities to align with additional cohort dictionaries. Figure 4 shows an excerpt of building the GECKO hierarchy by combining both CINECA specific terms in GECKO or reusing existing resources, and indicates classes natively in GECKO or imported for example from CMO.

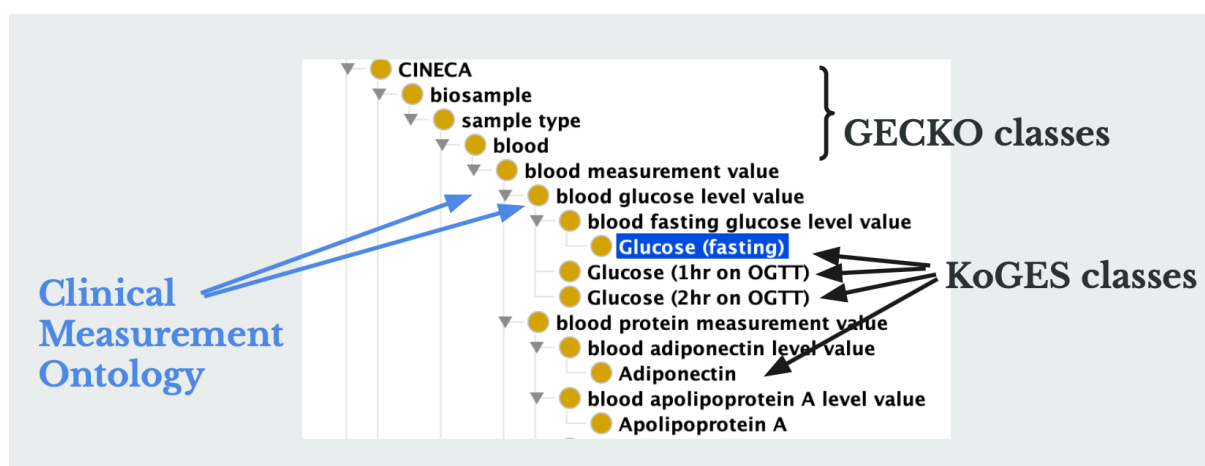


Figure 4. The GECKO includes new classes as well as those imported from reference ontologies such as CMO. Those are used to align with dictionaries from other cohorts, such as KoGES.

⁷ <https://www.maelstrom-research.org>

⁸ <https://creativecommons.org/licenses/by-nd/2.0/>

⁹ <http://www.obofoundry.org/>

¹⁰ <https://www.ebi.ac.uk/ols/ontologies/cmo>

¹¹ <http://obi-ontology.org/>

3.3.2 GECKO content

As of May 2021, the GECKO ontology contains 154 terms. 67 of the terms are in the GECKO namespace and 87 of them are imported from 15 other OBO Foundry ontologies shown in [Appendix 8.2](#), Table 1. Terms in GECKO have labels and stable, unique IDs. They also have textual definitions and optional additional annotation properties such as synonyms as shown in Figure 5.

The screenshot displays the GECKO ontology interface for the term "blood". At the top, the breadcrumb "OLS / Genomics Cohorts Knowledge Ontology" is followed by "GECKO" and "UBERON:0000178". A search bar labeled "Search GECKO" is on the right. Below the breadcrumb, the term "blood" is displayed with its URL "http://purl.obolibrary.org/obo/UBERON_0000178". A definition states: "A fluid that is composed of blood plasma and erythrocytes." Synonyms listed are "portion of blood" and "vertebrate blood". The left sidebar shows a tree view with "material entity" as the parent, followed by "organism substance", and "blood" as the selected term. A "Graph view" button is also present. The right sidebar, titled "Term information", shows the "IHCC browser label" as "blood", the "IHCC category" as "sample type", and the "has related synonym" as "whole blood". The "Term relations" section shows "Subclass of:" with "organism substance" as the only relation.

Figure 5. Term “blood” in the GECKO model.

GECKO contains six major categories identified in the minimal metadata model: biosample, basic cohort attributes, laboratory measures, questionnaire/survey data, statistics, and survey administration. In the ontological view of GECKO, all classes are organised under appropriate top-level classes from external ontologies such as the Basic Formal Ontology (BFO)¹² and the Information Artifact Ontology (IAO)¹³, so that GECKO can easily be aligned with other OBO Foundry ontologies. We also produce a cohort-friendly view of GECKO that organises the classes as shown in Figure 6 with the six major categories as the top-level. This view is better suited to browsing by cohort owners who understand those categories, versus the traditional ontology view which uses classes from the BFO as the top level, and underpins the faceted display in the cohort browser.

¹² <http://basic-formal-ontology.org>

¹³ <https://github.com/information-artifact-ontology/IAO/>



Figure 6. High level overview of the content of the GECKO as shown in the IHCC view. Each leaf category is further extended (not represented) to provide deeper granularity.

3.3.3 GECKO hosting

Zooma is a tool for annotating free text with ontology terms based on a curated repository of annotation knowledge. [Section 3.2.2](#) provides more details about its usage for cohort metadata harmonisation.

Finally, the Ontology Cross-references tool (OxO) contains cross references or mapping between terms from ontologies, vocabularies and coding standards. The default OxO service provided by the EMBL-EBI has data from a variety of sources including OLS and UMLS (Unified Medical Language System). We have repurposed OxO to work as a cross reference tool between cohort dictionaries and ontologies.

We deployed a containerized version of OLS alongside OXO and ZOOMA in the EBI Embassy cloud¹⁴. Cloud deployment and containerization reduces the management overhead by simplifying the maintenance tasks such as redeploying/reindexing after an addition of a new term. This registry¹⁵ hosts the latest version of the cohort dictionaries and supports search across them as well as export of the data using standardised APIs. Furthermore, OXO provides visualisation and data exploration tools for cohort dictionaries and ontologies. Figure 7 depicts a visualisation of mappings between GECKO and cohort dictionaries, as seen on the IHCC OxO instance. Figure 8 shows the number of mappings established.

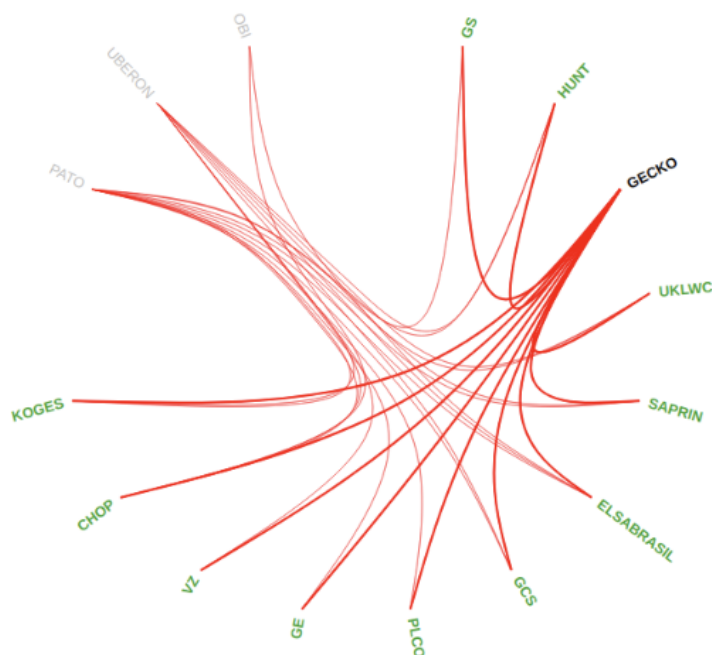


Figure 7. Mappings between ontology resources and cohort dictionaries as shown in the OxO tool.

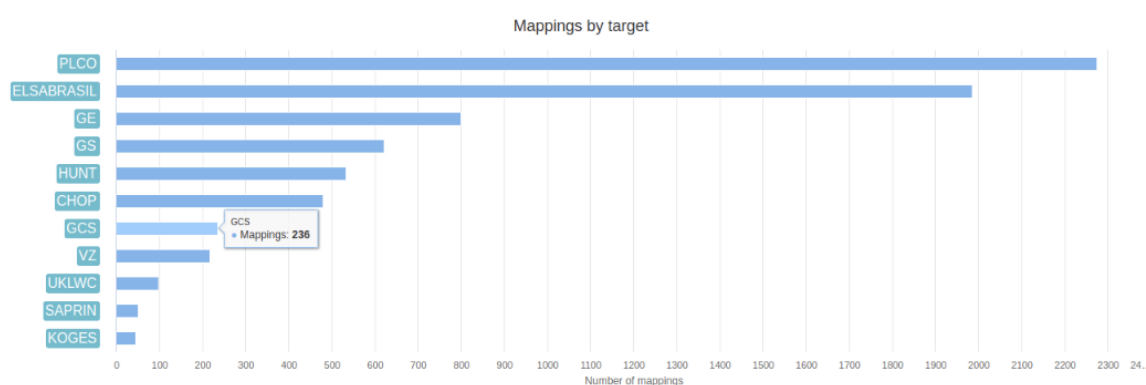


Figure 8. Number of mappings established across resources in the OxO tool.

¹⁴ <https://www.embassycloud.org/>

¹⁵ <https://registry.ihccglobal.app/ontologies>

3.4 Results

3.4.1 Federated cohort discovery in CINECA

Mapping CINECA synthetic cohort data to GECKO

The mapping between UK Biobank (UKB) data showcase attributes and the GECKO model was a manual undertaking, in which a set of intersecting attributes between two models was selected to generate synthetic data harmonised against the GECKO standard. UKB synthetic data covers all the main categories of the GECKO model and more than 90% of the sub-categories. The mapping process was not always one to one as there were many cohort fields that would map into a single term in the GECKO model. As an example, 'ever smoked', 'current tobacco smoking' and 'past tobacco smoking' would map into the single GECKO term of 'tobacco use history' (GECKO:0000068). Synthetic data for the UKB were initially derived using the TOFU¹⁶ tool, which produces randomly generated values based on the UKB data dictionary. Categorical values were randomly generated based on the data dictionary, continuous variables generated based on the distribution of values reported by the UKB showcase, and date / time values were randomly assigned. Additionally we split the phenotypes and attributes into four main classes - general, cancer, diabetes mellitus, and cardiac - based on common disease categories and use cases. We assigned the general attributes to all the samples, and the cardiac / diabetes mellitus / cancer attributes to a proportion of the total samples. Once the initial set of phenotypes and attributes were generated, the data was checked for consistency and where possible dependent attributes were calculated from the independent variables generated by TOFU. For example, BMI was calculated from height and weight data, and age at death generated by date of death and date of birth. These data were then loaded into the development instance of Biosamples¹⁷ which accessioned each of the samples.

Federated discovery of cohort data

The GECKO harmonised synthetic datasets were used for a demonstration presented at the CINECA March 2021 annual general assembly by WP1 - Federated Data Discovery and Querying. The datasets were loaded into GA4GH Beacon¹⁸ instances so that they could be queried programmatically by WP1. Results from simple questions such as "which cohorts have the data that is necessary for this use case?" as well as more complicated questions such as "in this dataset, what fraction of patients with disease X and variants in gene Y have outcome Z?" were returned. Loading the harmonised datasets into Beacon highlighted the need for further normalisation in addition to the current cohort harmonisation: values of the harmonised attributes also need to be standardised across cohorts. For example, while we can query for "tobacco history", values returned can be very diverse, e.g., "yes" "never" "3 packs a week" "0". This will be a focus of ongoing work in WP3. Another area of development will be extending GECKO beyond the minimal metadata model coverage - for example

¹⁶ <https://github.com/spiros/tofu>

¹⁷ <https://www.ebi.ac.uk/biosamples/>

¹⁸ <https://beacon-project.io>



adding granularity to “tobacco history”, such as “past smoker”, “never smoked” or “occasional current use”.

3.4.2 Federated cohort discovery beyond CINECA

To demonstrate the wider usefulness of GECKO and the applicability of our harmonisation best practices, we have built a prototype framework providing standardised methods for discovery and presentation of cohort harmonised metadata. Our proof-of-concept implementation demonstrates that we can harmonise a specific subset of metadata on a limited number of cohorts, and deploy a service allowing federated browsing on them and discovery. While adding a new cohort dictionary to the registry is inherently highly variable depending on the size and complexity of the dictionary, in our current work medium-size dictionaries require roughly one full day from cohort owners and technical support for addition.

The IHCC cohort atlas

The IHCC Cohort Atlas, shown in Figure 9, is a public catalogue of standardised IHCC cohorts for federated discovery and data sharing. It provides an opportunity to accelerate discovery and translation by sharing and integrating data from potentially over 50 million people across the world for the first time. The IHCC Cohort Atlas demonstrates federated discovery of harmonised cohorts and faceted filtering based on phenotypic, genomic, and exposure data among others. Additional screenshots of the browser and filters are provided in [Appendix 8.1](#).

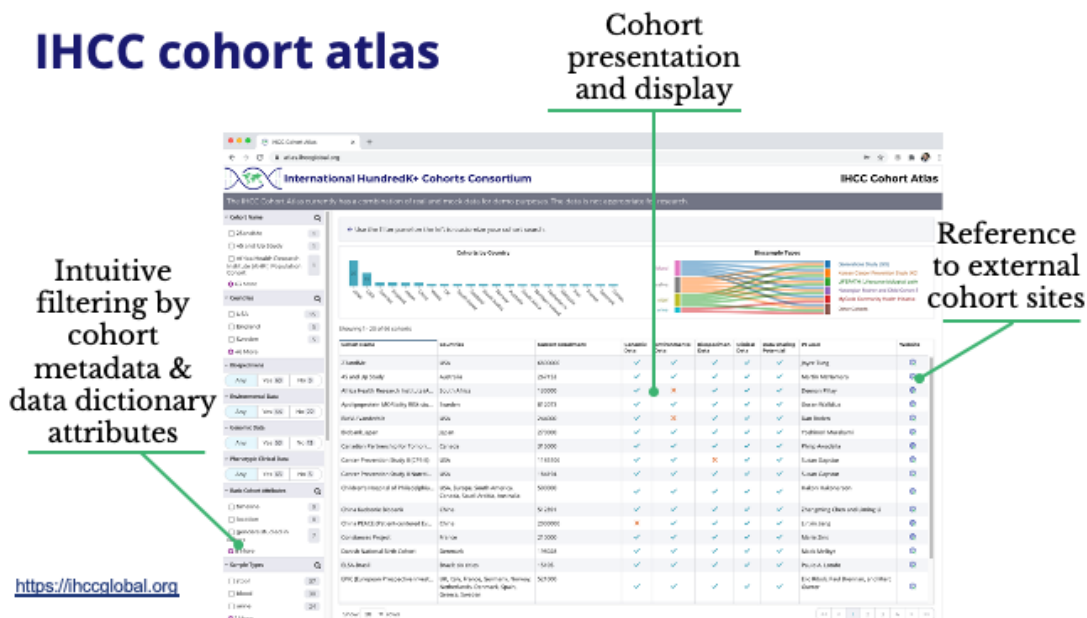


Figure 9. GECKO harmonised cohorts in the IHCC atlas. To illustrate the diversity of GECKO based searches this includes cohort data as well as pseudo-data.

The Davos Alzheimer's Collaborative Atlas

We have further reused our GECKO-powered platform in the Davos Alzheimer's Collaborative (DAC) project¹⁹, see Figure 10.

The DAC Atlas catalogues metadata from cohorts around the world, allowing for cohort 'discoverability' by Alzheimer's disease investigators. It is the first step to building an Alzheimer's global platform for translational research. Reusing the GECKO model in the DAC project demonstrated that the model is extensible as several additional attributes were added, such as imaging data types. It also showcased the advantage of using standards and shared infrastructure as we were able to quickly repurpose the work done under the IHCC atlas. The DAC has secured additional funding and will keep expanding the content of the Atlas, further leveraging and enhancing the GECKO powered infrastructure we deployed.

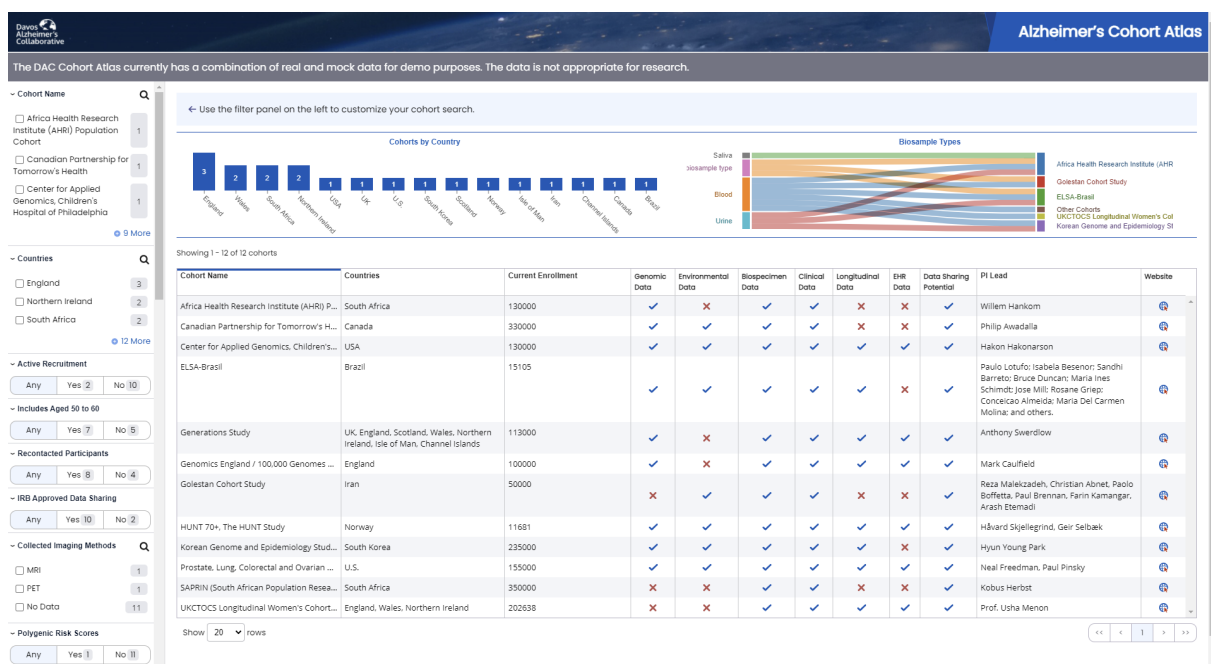


Figure 10. The Davos Alzheimer's Collaborative atlas. GECKO-based facets visible in the left hand side menu have been extended to include Alzheimer specific attributes such as imaging data.

3.5 Next steps

We have completed our tasks for this deliverable. In addition to providing a semantic metadata model based on the CINECA minimal metadata model and best practices for cohorts metadata harmonisation, we have interacted and engaged with external communities such as Maelstrom and OBO Foundry to increase interoperability, and IHCC and DAC to promote reuse. Indeed, the GECKO has already successfully been reused in at least two different implementations outside of CINECA to enable cohort discovery.

¹⁹ <https://www.davosalzheimerscollaborative.org>

3.5.1 Interoperability with Maelstrom

As described in [section 3.3.1](#), the original CC-BY-ND licence did not allow for reuse and extension of the Maelstrom taxonomy, requiring duplication of the Maelstrom content in GECKO to add more granularity to the existing categories as required by the CINECA use cases. After extensive discussion with the Maelstrom group, they have agreed both to update their licensing terms and on a schema for representing Maelstrom categories in OWL, and converted the terms to achieve this. We will replace existing GECKO classes with their original Maelstrom counterpart, thereby enabling native interoperability with projects already using the Maelstrom classes.

3.5.2 Engagement with other communities

We will extend the Phenopackets GA4GH standard²⁰ to include the GECKO attributes. This will make those attributes amenable to query via the upcoming GA4GH Beacon v2 network using a standard exchange format. In addition, the Phenopacket JSON schema representation in Beacon will also enable validation through the ELIXIR biovalidator²¹, to guarantee compliance of metadata captured to best practices when and if they evolve.

While the CINECA cohorts did not reuse any semantic framework but for Maelstrom categories, several models are available for metadata, such as the OMOP Common Metadata Model²², the Clinical Data Interchange Standard Consortium (CDISC) standards²³ and the HL7 Fast Healthcare Interoperability Resources (FHIR)²⁴. While not directly in the scope of this deliverable, when a cohort using any of those models will join the atlas browser the mappings between GECKO and the model will be captured and stored in our Oxo instance, making them directly reusable by other cohorts.

Finally, to improve interoperability with the wider community, and support cross portal querying, we have undertaken to merge the admin metadata model with other community models, including BBMRI-ERIC. The model compiled in JSON Schema facilitates validation of cohort admin data and ensures conformity between cohorts. The resulting cohort admin data will be used for developing a CINECA cohort portal, exporting/importing admin data with similar portals and will enable cross portal querying.

The GECKO source files, scripts and documentation are licensed under CC-BY 4.0 and available from the Github repository <https://github.com/IHCC-cohorts/GECKO>. GECKO has been included in the OBO Foundry, making it widely available to the scientific community.

²⁰ <https://github.com/phenopackets/phenopacket-schema/tree/cohort/src>

²¹ <https://www.npmjs.com/package/elixir-jsonschema-validator>

²² <https://www.ohdsi.org/data-standardization/the-common-data-model/>

²³ <https://www.cdisc.org>

²⁴ <https://www.hl7.org/fhir/>



This will provide the starting point for community-led development of a Minimum Information standard for cohorts, and will align existing standards such as MIABIS²⁵.

4. References

Lee, Meng-Tse Gabriel, Tzu-Chun Hsu, Shyr-Chyr Chen, Ya-Chin Lee, Po-Hsiu Kuo, Jenn-Hwai Yang, Hsiu-Hao Chang, and Chien-Chang Lee. 2019. "Integrative Genome-Wide Association Studies of eQTL and GWAS Data for Gout Disease Susceptibility." *Scientific Reports* 9 (1): 4981.

5. Abbreviations

DAC	Davos Alzheimer's Collaborative
EuCan	European and Canadian consortium projects
GA4GH	The Global Alliance for Genomics and Health
GECKO	Genomics Cohorts Knowledge Ontology
IHCC	International HundredK+ Cohorts Consortium
JSON	JavaScript Object Notation
OLS	Ontology Look-up Service
OWL	Web Ontology Language
REST API	Representational State Transfer Application Programming Interface
WP	Work Package

6. 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

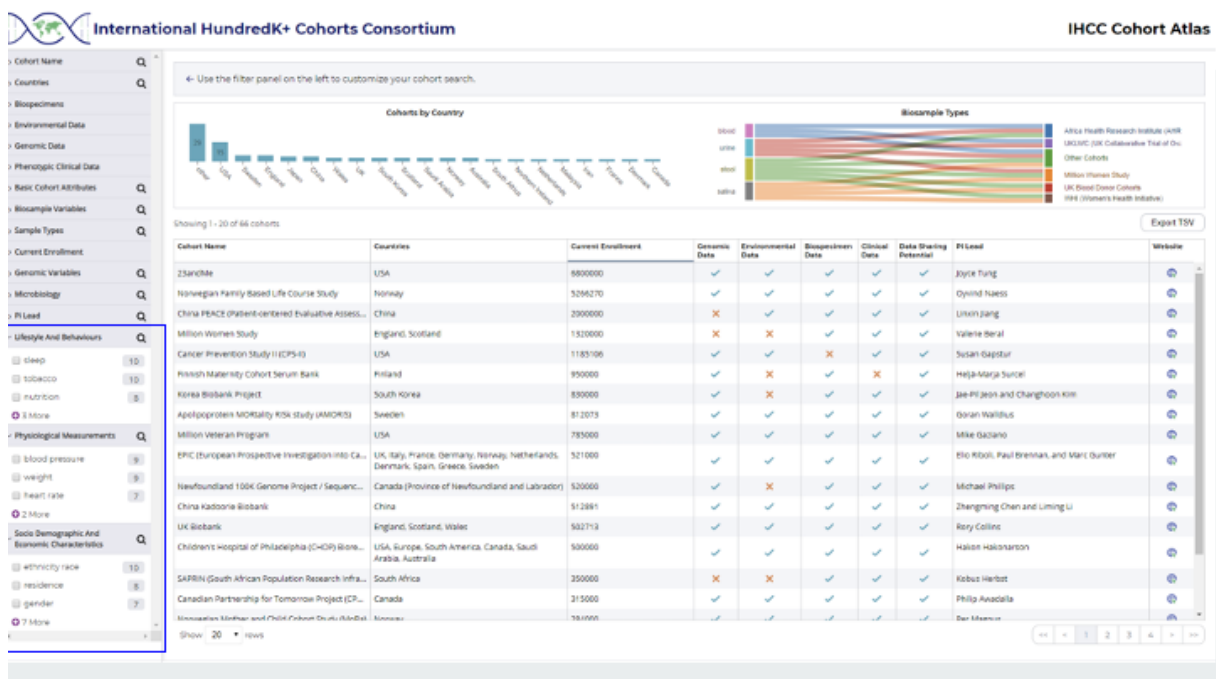
7. Delivery and schedule

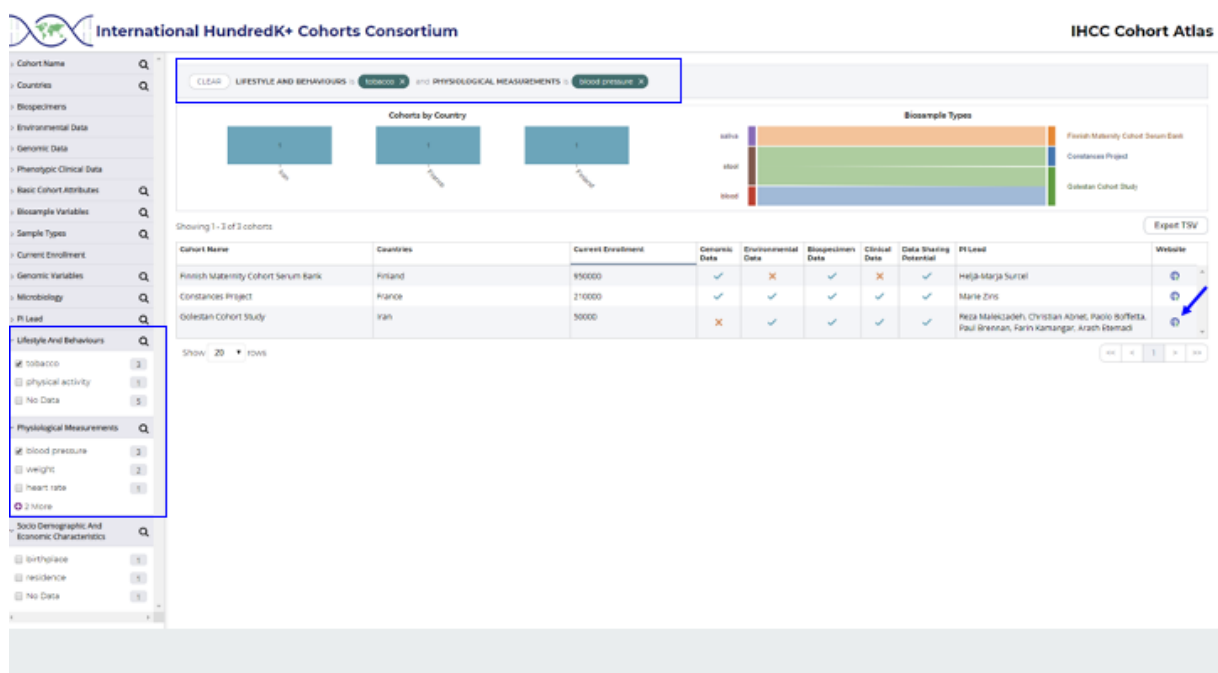
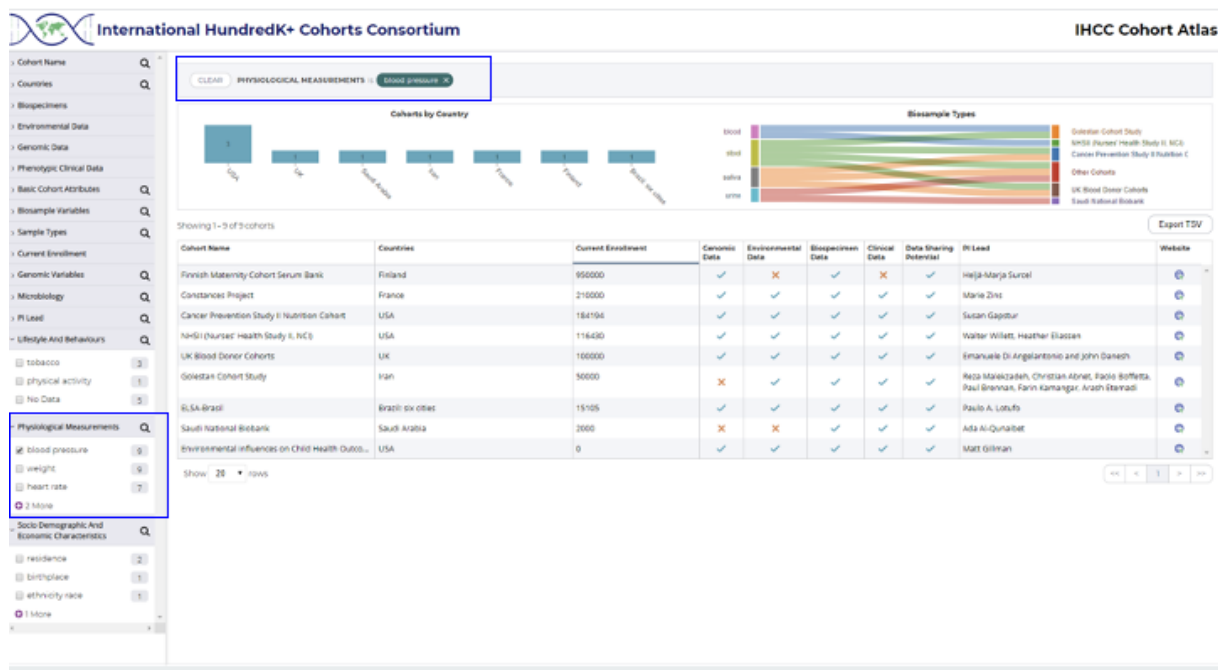
The delivery is on time.

²⁵ <https://github.com/MIABIS/miabis/wiki>



8.1 The IHCC Cohort Atlas: faceted browsing and discovery using GECKO





NIH NATIONAL CANCER INSTITUTE
Division of Cancer Epidemiology & Genetics

GEMINI Shared Repository (GEMShare)

Home About Requests Studies FAQ

Home / Studies / Golestan Cohort Study (GCS)

[Request Study Data and/or Samples](#)

Golestan Cohort Study (GCS)

This cohort study is evaluating the environmental and genetic risk factors for esophageal squamous cell carcinoma (ESCC) in Golestan Province, Iran. The study is led by the Digestive Diseases Research Institute of Tehran University of Medical Sciences in collaboration with NCI Division of Cancer Epidemiology and Genetics, and the International Agency for Research on Cancer (IARC) in Lyon. From 2004 – 2008, the study recruited approximately 50,000 adults. The study population is a sample of the Golestan population, aged 40-75 years. Baseline assessments included a lifestyle questionnaire, a semi-quantitative food frequency questionnaire, and collection of blood, hair, nails and urine. Measurements included height, weight, waist and hip circumference, body size at different ages, and physical activity. Annual follow-up is ongoing.

Study Publications (0)

Documents

No documents available

For more information contact:

Arash Etemadi (arashetemadi@nih.gov), Hossein Poustchi (h.poustchi@gmail.com)

Contact Us Policies Accessibility FOIA

U.S. Department of Health and Human Services National Institutes of Health National Cancer Institute USA.gov

NIH... Turning Discovery Into Health®

International HundredK+ Cohorts Consortium

IHCC Cohort Atlas

Coort Name Countries Biospecimens Environmental Data Genomic Data Phenotypic Clinical Data Basic Cohort Attributes Biosample Variables Sample Types Current Enrollment Genomic Variables Microbiology PI Lead Lifestyle And Behaviours

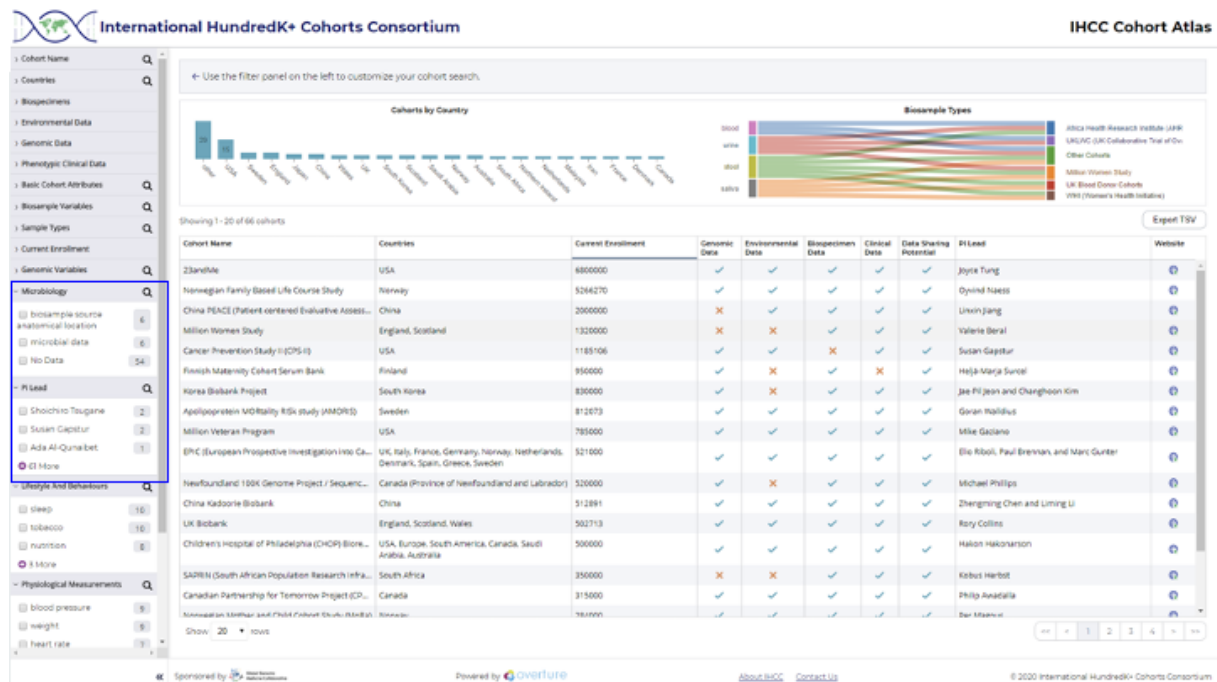
Physical Measurements

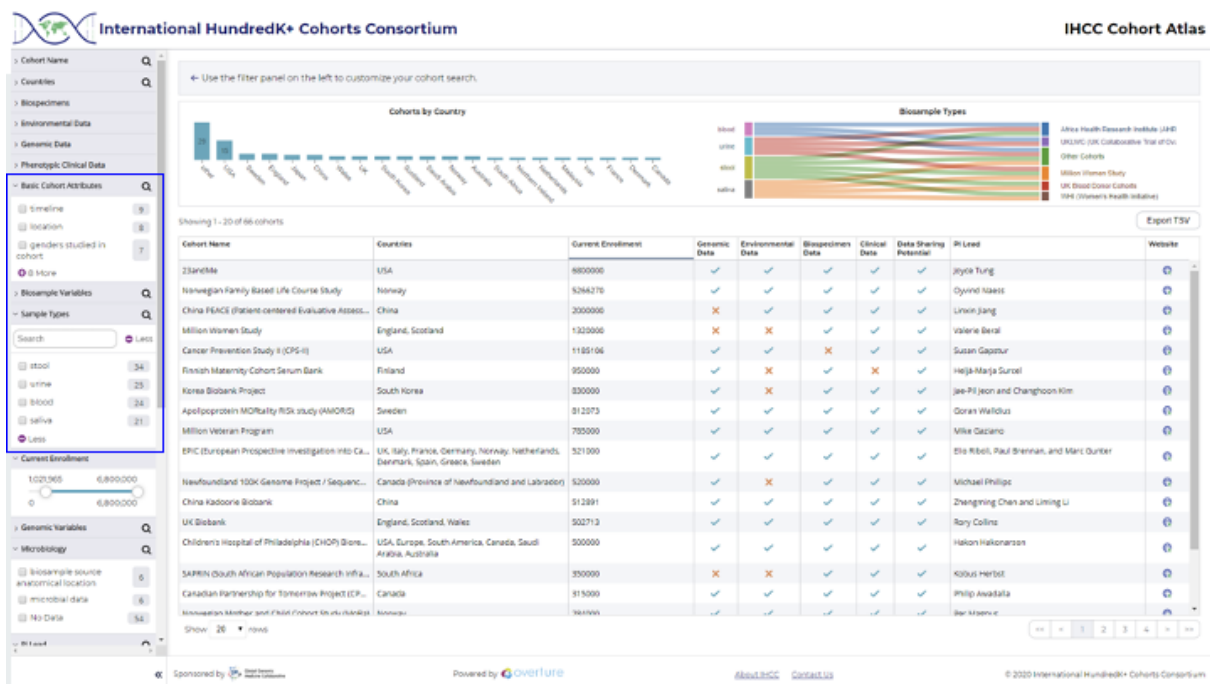
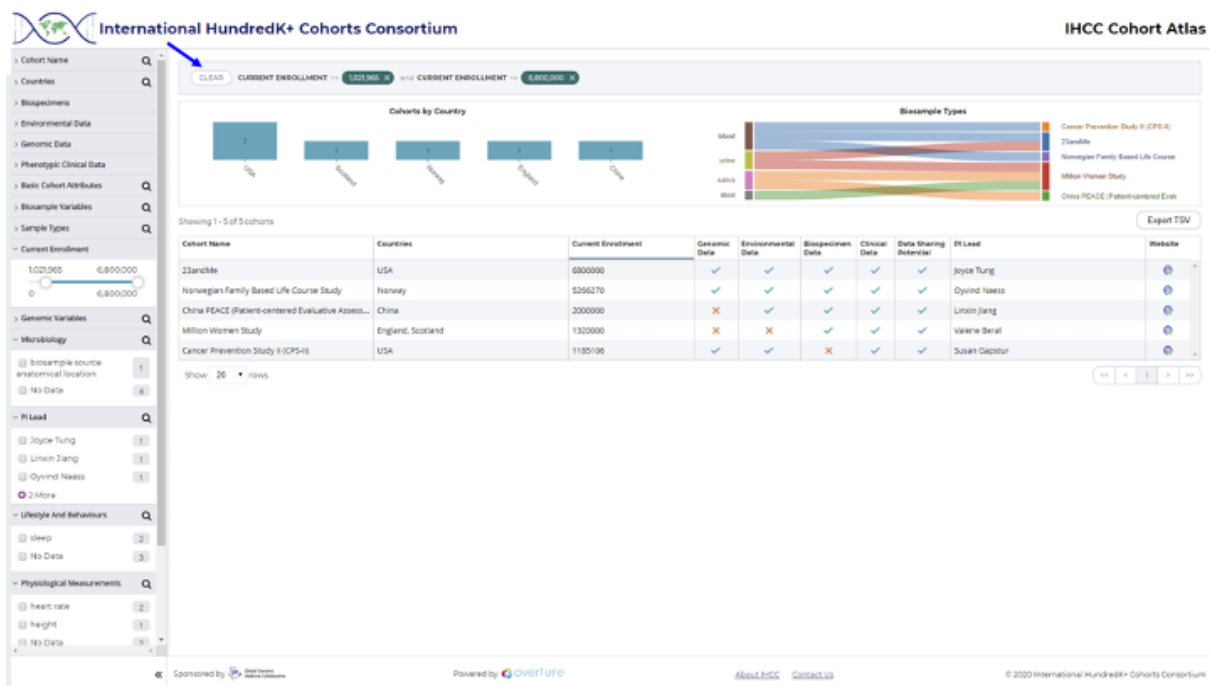
Birthplace Residence No Data

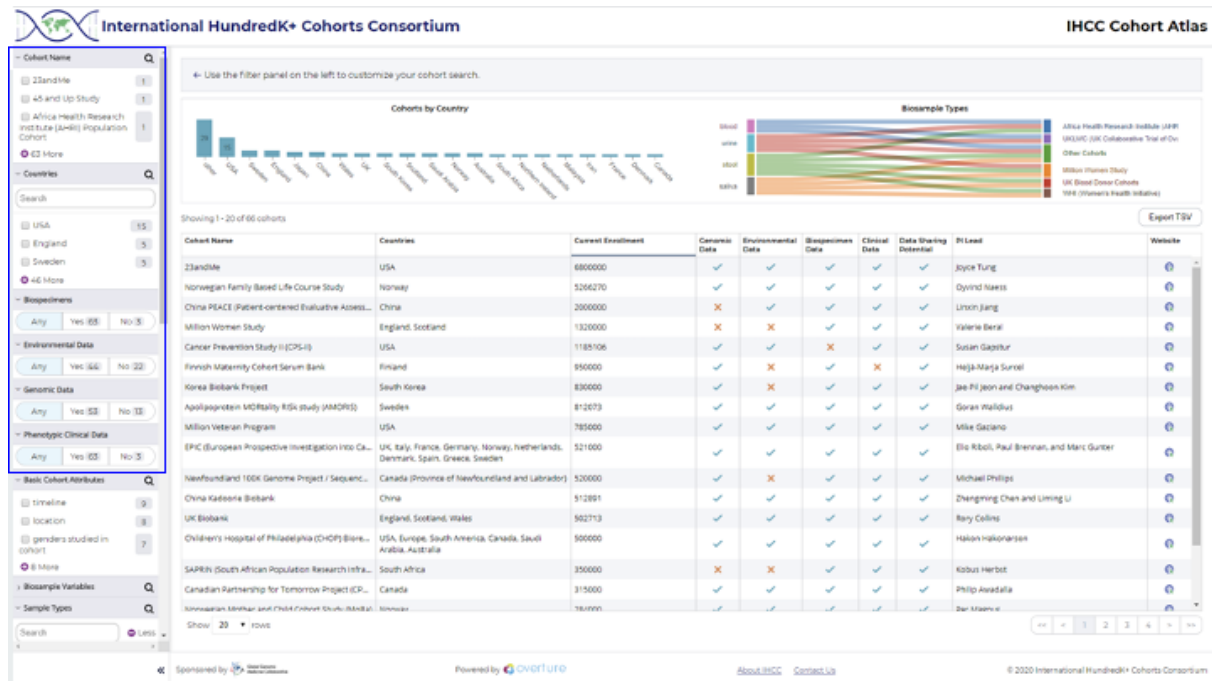
Showing 1 - 3 of 3 cohorts

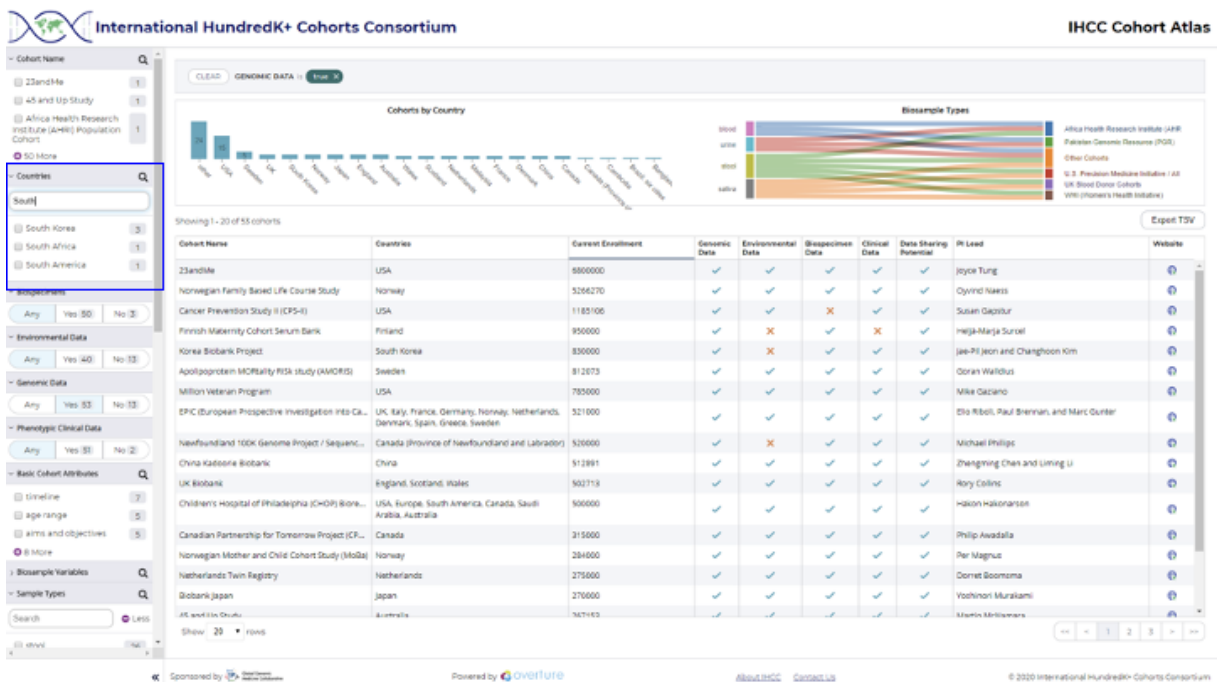
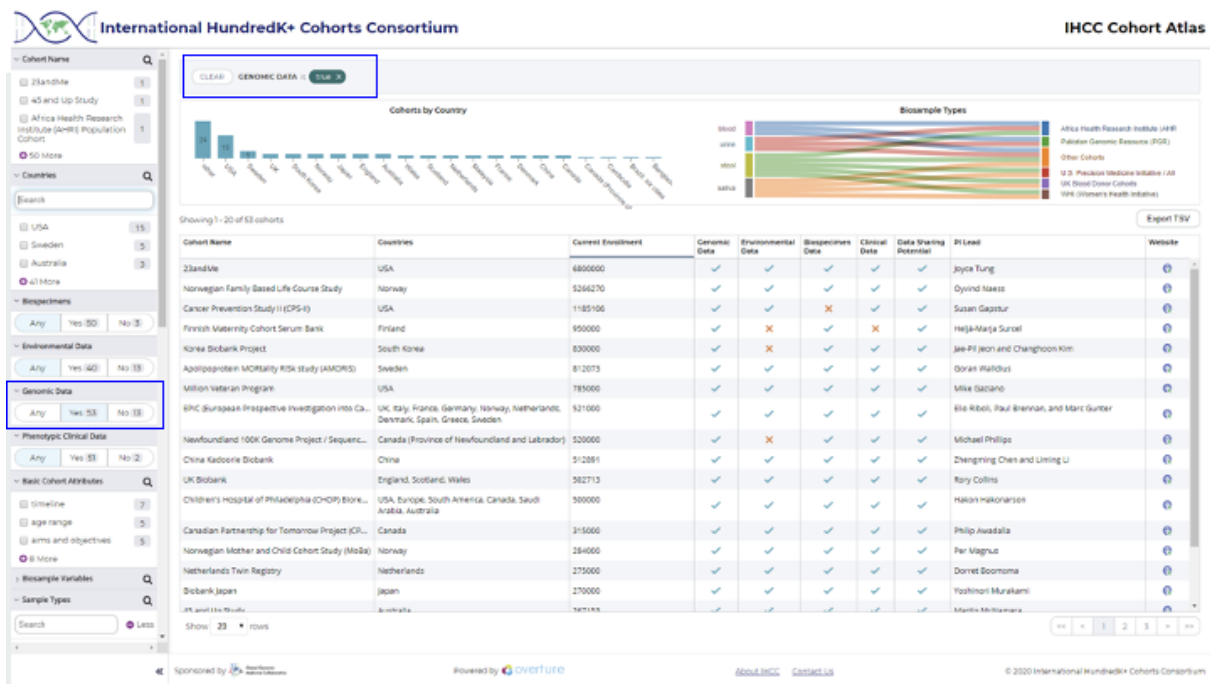
Cohort Name	Countries	Current Enrollment	Demographic Data	Environmental Data	Biospecimen Data	Clinical Data	Data Sharing Potential	PI Lead	Website
Finnish Maternity Cohort Serum Bank	Finland	90000	✓	✗	✓	✗	✓	Heli-Järvelin, Sirkka	Website
Constances Project	France	210000	✓	✓	✓	✓	✓	Marie Zins	Website
Golestan Cohort Study	Iran	50000	✗	✓	✓	✓	✓	Peziz Malekzadeh, Christian Adami, Paolo Buffeta, Paul Brennan, Farshad Kamangar, Arash Etemadi	Website

Show: 20 rows









8.2 Resources imported in GECKO

Resource Name	Prefix	Terms in GECKO
Basic Formal Ontology	BFO	7
Clinical Measurement Ontology	CMO	7
VEuPathDB Ontology	EUPATH	7
Gene Ontology	GO	1
Information Artifact Ontology	IAO	11
Mental Functioning Ontology	MF	1
Mondo Disease Ontology	MONDO	19
Ontology for Biomedical Investigations	OBI	16
Ontology for General Medical Science	OGMS	2
Ontology of Medically Related Social Entities	OMRSE	1
Phenotype and Trait Ontology	PATO	3
Population and Community Ontology	PCO	1
Prescription of Drugs Ontology	PDRO	2
Statistical Methods Ontology	STATO	3
Uberon Multi-species Anatomy Ontology	UBERON	6

Table 1. Ontologies reused by GECKO and the corresponding number of terms. Not all of these imported terms are used for alignment; some are for upper-level organisation of GECKO, such as those from BFO.

