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

Deliverable D1.1
Discovery Service Catalogue

Work Package:	WP1 - Federated Data Discovery and Querying
Lead Beneficiary:	European Molecular Biology Laboratory
WP Leader:	Jonathan Dursi (SickKids)
Contributing Partner(s):	SickKids, CRG, HES-SO, UCT
Contractual Delivery Date:	30th June, 2020
Actual Delivery Date:	23rd June, 2020
Authors of this Deliverable:	Jonathan Dursi (SickKids) Jordi Rambla de Argila (CRG) Sabela de la Torre (CRG) Romain Tanzer (HES-SO) Nona Naderi (HES-SO) Mamana Mbiyavanga (UCT) Samarth Agarwal (SickKids)
Reviewed by:	Mikael Linden (CSC) Kaur Alasoo (Tartu)
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 the federated queries and analyses of distributed cohorts across continents. A vital component of this work is building a machine readable catalogue of cohorts and sites that support the efforts of Work Package 1 (WP1) discovery and analysis APIs, which can be programmatically queried so that API calls can be made to relevant sites and results gathered and presented to the researcher.

Deliverable D1.1, Discovery Service Catalogue, supports the work of and dependent work packages by implementing and demonstrating an open-source extended implementation of the Service Registry standard[1] of the Global Alliance for Genomics and Health (GA4GH) for WP1's discovery queries, the GA4GH Beacon[2] queries. The Service Registry standard is now supported by the ELIXIR Beacon Network[3] that CINECA WP1 uses to federate discovery queries across cohorts, and this demonstrator deliverable demonstrates the use of the service registry and its open source implementation.

2. Project objectives

This deliverable has contributed to the following objectives:

a) Demonstrating the suitability of the ELIXIR-extended GA4GH Service Registry standard for listing sites and cohorts responding to WP1 queries.
b) Implementing a first service catalogue of CINECA APIs for performing discovery queries.
c) Presenting an open-source implementation that can be used for analysis queries and other CINECA APIs.

3. Detailed report on the deliverable

3.1 Background

For researchers to be able to make discovery or analysis queries across multiple cohorts and sites, there needs to be a list of such cohorts and sites available. This should be programmatically readable, so that updates are promptly reflected in new queries, and should be implemented in such a way that multiple tools can make use of it (e.g., multiple portals, command line tools, etc).

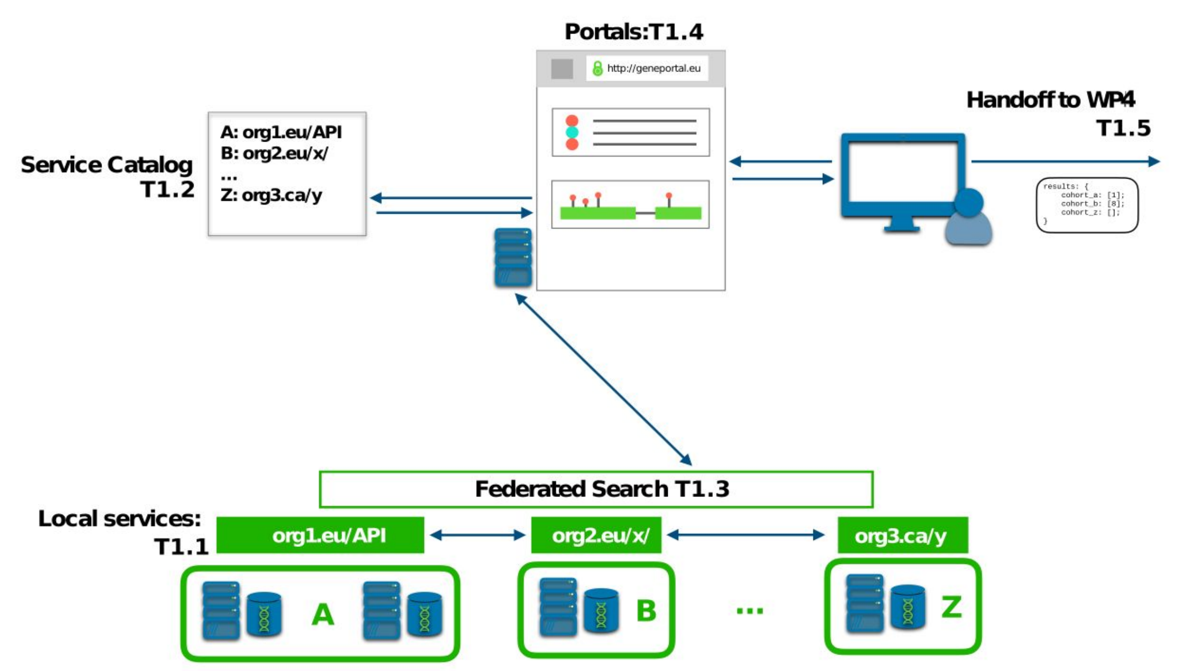
To ensure that the work of CINECA has the maximum impact, the work packages wherever possible make use of standards - either existing standards, or working with standards organisations to ensure that emerging standards meet the needs of the CINECA researcher community. Using standards-based approaches improves the FAIRness[4] of what we are building by ensuring Interoperability between our services and that of the larger health informatics ecosystem.

Table 1, below, shows the components that WP1 is delivering to the CINECA platform.

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

Service	Purpose	Questions Answered
Basic Queries	Discovery of relevant data and cases	Does this dataset have any patients with disease X and genotype data?
Extended Queries	Interactive analyses	In this dataset, what fraction of patients with disease X and variants in gene Y have outcome Z?
Service Registry	Catalogue of services	What data sets are out there and what queries can I make of them?
Portals	Interactive exploration	Examining data with all queries, CINECA-wide

These services interact with each other to provide a consistent platform for interactive queries as shown in Figure 1. Basic (Discovery) and Extended (Analysis) queries are deployed at CINECA sites to make CINECA cohort data available for querying and analysis. The service catalogue provides an authoritative listing of the services and cohorts available, to support unified portals and federated access.



```
results: {
    cohort_a: [1];
    cohort_b: [8];
    cohort_z: [];
}
```

Figure 1: The architecture of WP1 APIs and tasks for supporting analysis and discovery queries. The service catalogue is a key part of federating the queries across sites and cohorts.

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

The GA4GH has this year released a service registry standard[1], a read-only API that allows users to query lists of GA4GH (or other) APIs that are known with the service registry. This information is stored and reported in a standardised manner defined in another standard, the Service Info standard[5].

WP1 has decided on the use of the GA4GH Beacon[2] for discovery queries, and the ELIXIR Beacon Network[3] as a mechanism for federating those queries across multiple sites. Work by CINECA partner CRG has gone into implementing the service registry into the Beacon Network. Registering beacon APIs at various sites into the Beacon Network means actually, behind the scenes, populating a Service Registry which can be queried by other tools such as portals or command line tools.

3.2 Work Done

For this demonstration, CINECA partners at the CRG have implemented the GA4GH service registry[1] standard leveraging a previous version of the Registry, designed initially in the ELIXIR Beacon Network context, and an evolution of the Service Info[5] standard in the reference Beacon v2.x implementation[6]. The standard has also been extended to allow basic information about the cohort data available through the service to be included in the response. This allows clients using the service registry to evaluate which services, and so which cohorts, to query.

The demonstration consists of three pre-recorded webinars. First is an introductory “explainer” video, describing why we need the Service Registry and the role it plays in the larger efforts in WP1. Second, we have a demonstration of how a standalone service registry operates. Third, we have a screen-shared technical demonstration, showing the use of the service registry.

The technical demonstration consists of seven parts:

1. A demonstration of a Web UI for the Service Registry
2. Showing how to list the registered Services through the Service Registry API
3. Showing a Service-info response from a given service, which is where the information in the Service Registry comes from
4. Showing how a new Service is added to the Service Registry
5. Showing the newly added Service in the Service Registry
6. Listing the cohorts available through the Services
7. Post a Discovery (Beacon) query in the Service Registry and get the response back

Videos of the recorded webinars are available as follows:

- An introduction to CINECA’s Service Catalog - Deliverable 1.1
  Summary: This video introduces the Service Catalog, a searchable listing of query services available atop CINECA cohort data, and explains where it fits into CINECA’s set of interactive queries and data federation. Length of the video: 3:20.
  Published on the CINECA YouTube channel here.

- CINECA Discovery Service Catalog - Technical Walkthrough - Standalone Service

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

Summary: This video demonstrates how a standalone service registry operates, and our extensions to the service info and service registry standards to include cohort-level metadata.
Length of the video: 4:36.
Published on the CINECA YouTube channel here.

- CINECA Discovery Service Catalog - Technical Walkthrough - Beacon Integration
  Summary: This video demonstrates how a service registry can operate with additional functionality when integrated more closely with the services providing queries (here, Beacon queries). Length of the video: 6:38.
  Published on the CINECA YouTube channel here.

3.3 Next steps

With the Discovery Service Catalogue now implemented and demonstrated, CINECA WP1 (federated discovery and analysis queries) will proceed to implementing federated discovery queries across cohorts for Milestone 1.2, due in December (M24). We aim to have that milestone completed ahead of schedule for the mid-term review.

In addition, WP1 is working on implementing our extended queries with the emerging Search standard by GA4GH. As this service is deployed, we will list those services in the CINECA-wide service registry as well.

On the policy side, WP1 will work with project management, the cohorts, and sites to determine the proper level of authentication and authorisation that should be required to access the service registry. We will then work with WP2 to implement that authentication and authorisation policy in the service registry.

Finally, to ensure that the work of CINECA is as broadly shared as possible, the extension to service info and service registry standards will be proposed to the GA4GH. This will allow other federated cohort projects to benefit from the work done by CRG and CINECA.

4. References

[1] Service Registry. Service Registry API v1.0 (2020). Available at:
https://github.com/ga4gh-discovery/ga4gh-service-registry. (Accessed: 17th May 2020)

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

[3] Fiume, M., Cupak, M., Keenan, S. et al. Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol* 37, 220–224 (2019). https://doi.org/10.1038/s41587-019-0046-x

[4] GO FAIR. FAIR principles (2016). Available at https://www.go-fair.org/fair-principles/. (Accessed: 17 May)

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

[5] Service Info. GA4GH Service-Info Specification v1.0 (2020). Available at: https://github.com/ga4gh-discovery/ga4gh-service-info. (Accessed: 17th May 2020)

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

5. 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

6. Delivery and schedule

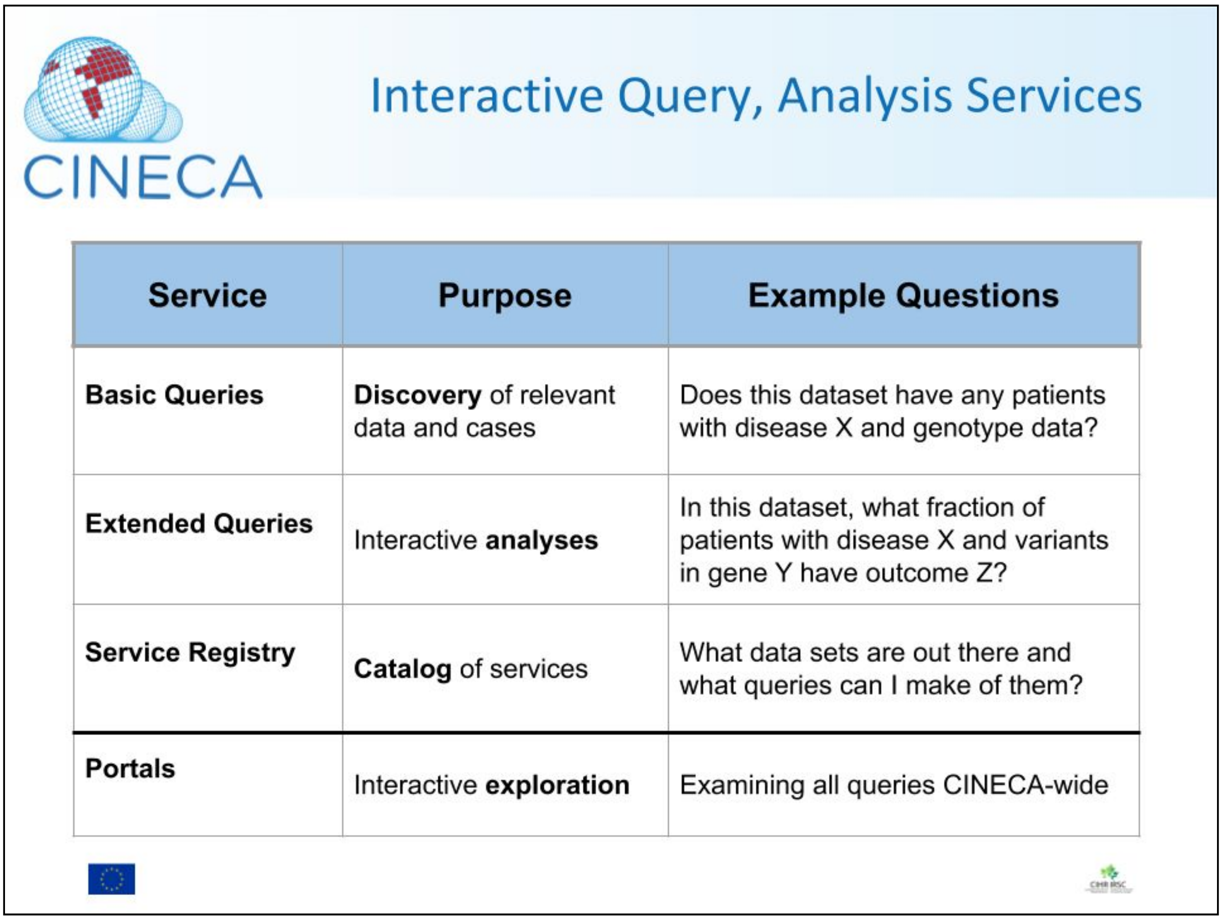
The delivery is on time.

7. Appendices

7.1 Appendix 1 - A Catalogue For CINECA Services, Introduction

This appendix includes a summary of the slides presented in the introductory video for the Discovery Service catalogue.

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

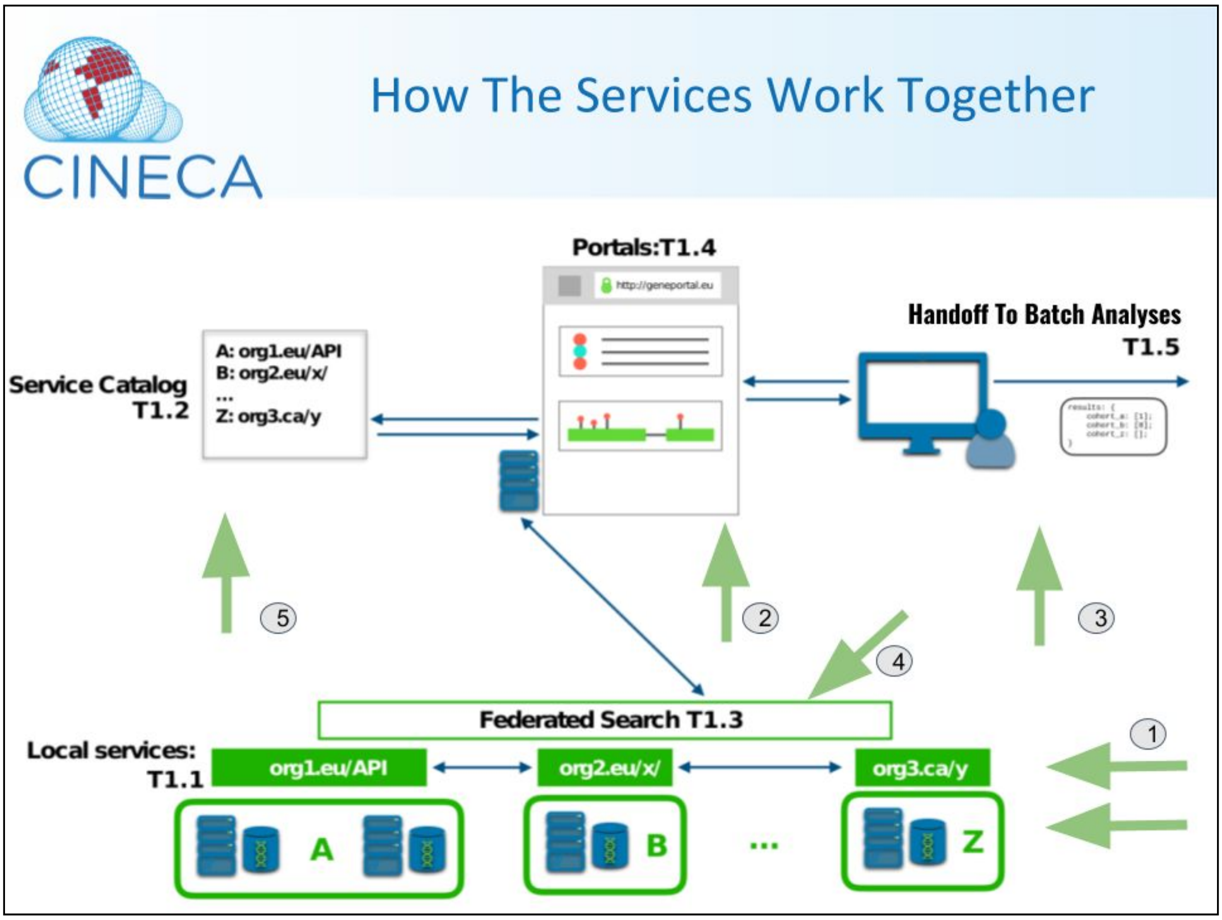


Interactive Query, Analysis Services

Service	Purpose	Example Questions
Basic Queries	**Discovery** of relevant data and cases	Does this dataset have any patients with disease X and genotype data?
Extended Queries	Interactive **analyses**	In this dataset, what fraction of patients with disease X and variants in gene Y have outcome Z?
Service Registry	**Catalog** of services	What data sets are out there and what queries can I make of them?
Portals	Interactive **exploration**	Examining all queries CINECA-wide

The goal of our team is to allow CINECA researchers to interactively discover, query, and analyze data from across CINECA cohorts. An authorised researcher should be able to transparently make uniform queries across the CINECA federation of cohort data and get back the results they need. To do this, we are building and deploying four separate sets of tools.

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



How The Services Work Together

```
results: {
    cohort_a: [1];
    cohort_b: [0];
    cohort_z: [];
}
```

This diagram shows how the pieces our team are building fit together.

1 - Basic and extended query services are being implemented at the sites.

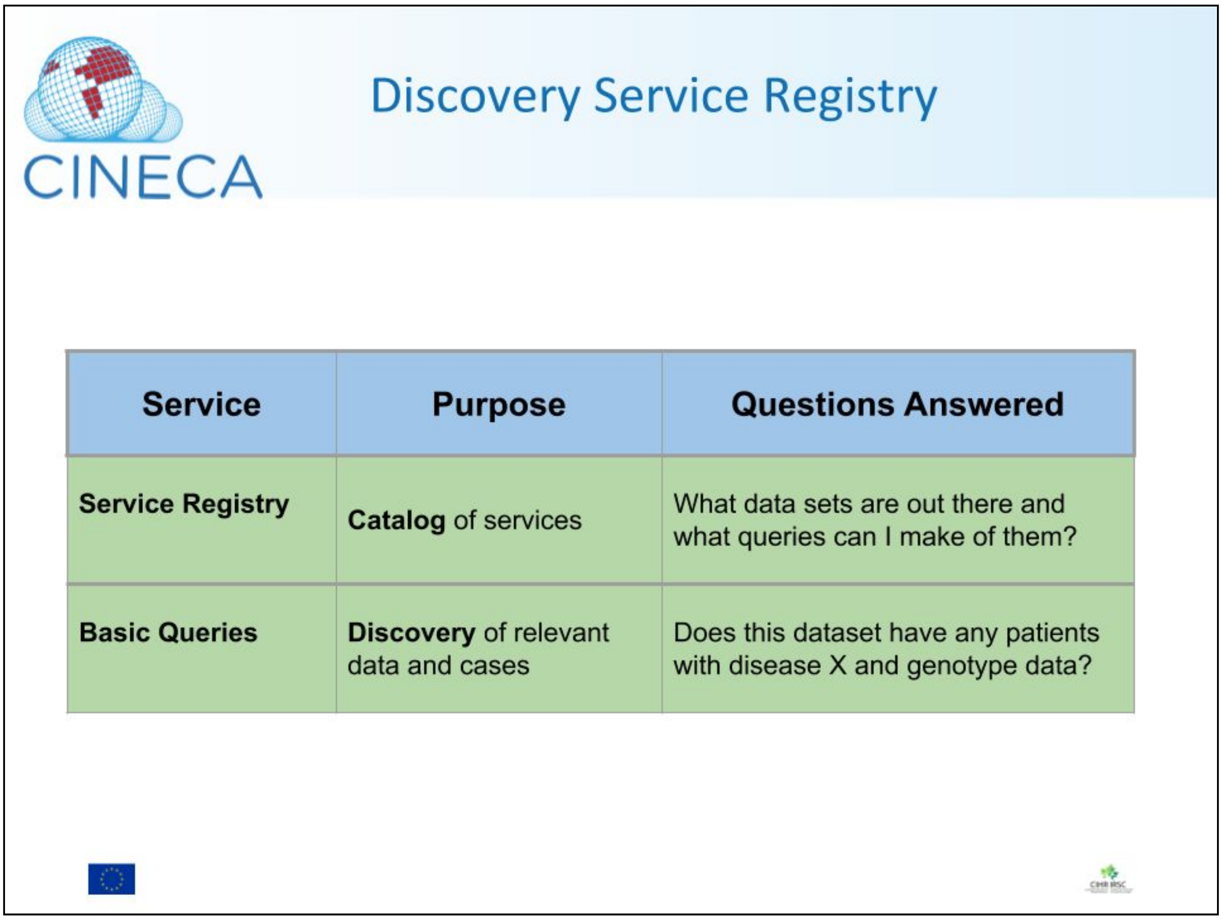
2 - The portal will appear later, providing results directly to researchers.

3 - Also providing results to the services of other teams for long-running analyses.

4 - This enables the federated search across cohorts, including the portal.

5 - This is the service registry we are developing.

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

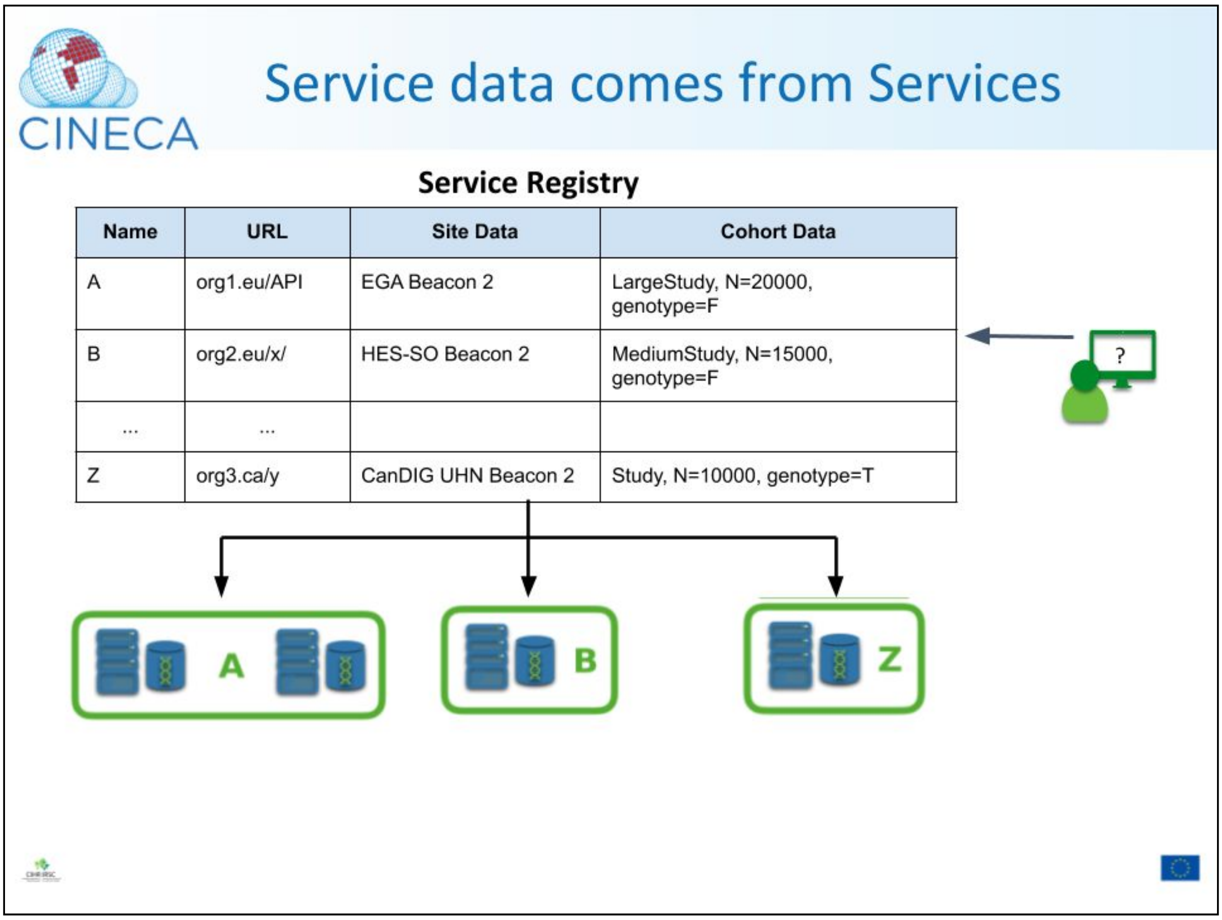


Discovery Service Registry

Service	Purpose	Questions Answered
Service Registry	**Catalog** of services	What data sets are out there and what queries can I make of them?
Basic Queries	**Discovery** of relevant data and cases	Does this dataset have any patients with disease X and genotype data?

Our demonstration of the service registry includes services that support our Basic Queries, which we have implemented with the GA4GH Beacon standard, as implemented and extended by ELIXIR and the CRG.

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



CINECA

Service data comes from Services

Service Registry

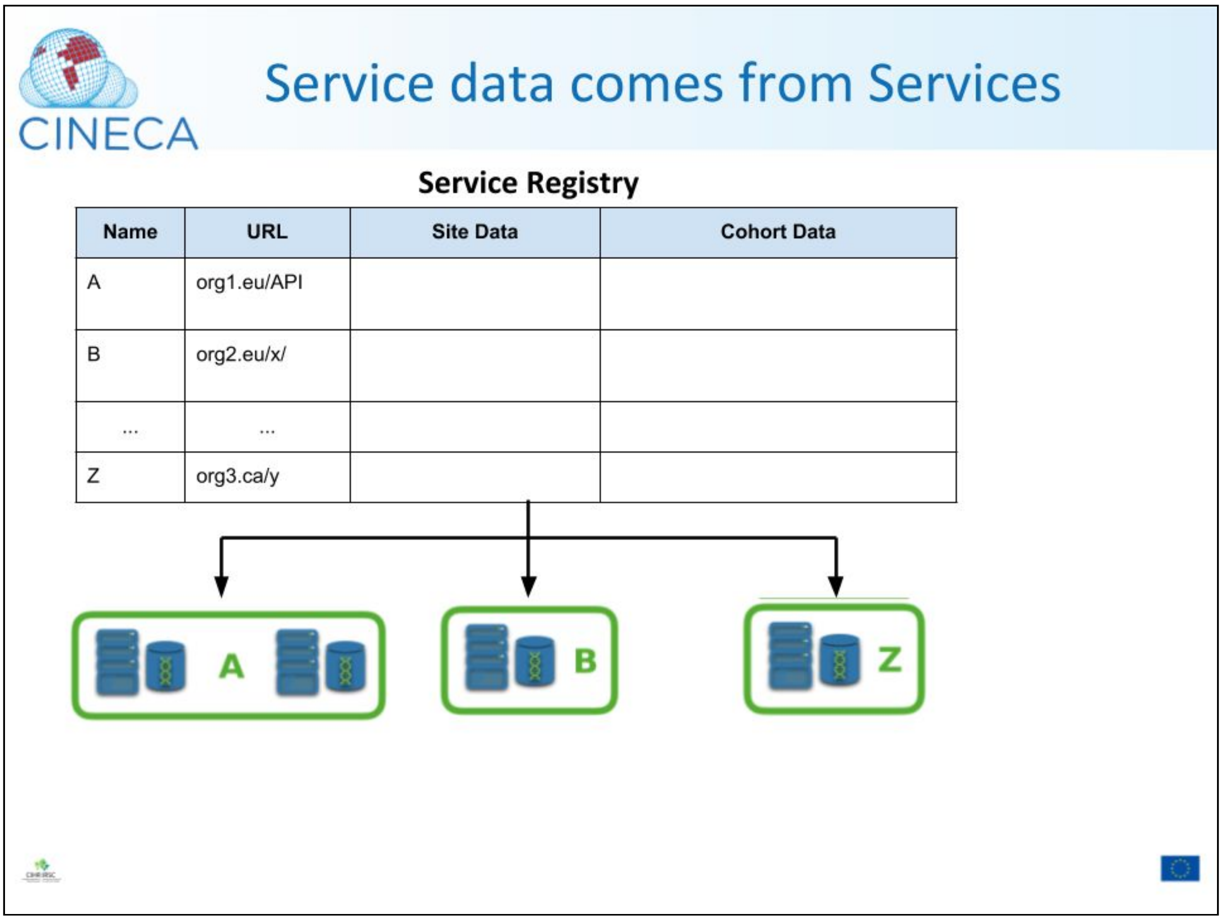
Name	URL	Site Data	Cohort Data
A	org1.eu/API	EGA Beacon 2	LargeStudy, N=20000, genotype=F
B	org2.eu/x/	HES-SO Beacon 2	MediumStudy, N=15000, genotype=F
...	...		
Z	org3.ca/y	CanDIG UHN Beacon 2	Study, N=10000, genotype=T

The service registry can be queried directly, and also used to pass through queries to the individual services. The second appendix will show the details.

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

7.2 Appendix 2 - A Catalogue For CINECA Services, Technical Walkthrough

This appendix includes a summary of the slides presented in the technical walkthrough demo video for the Discovery Service catalogue.



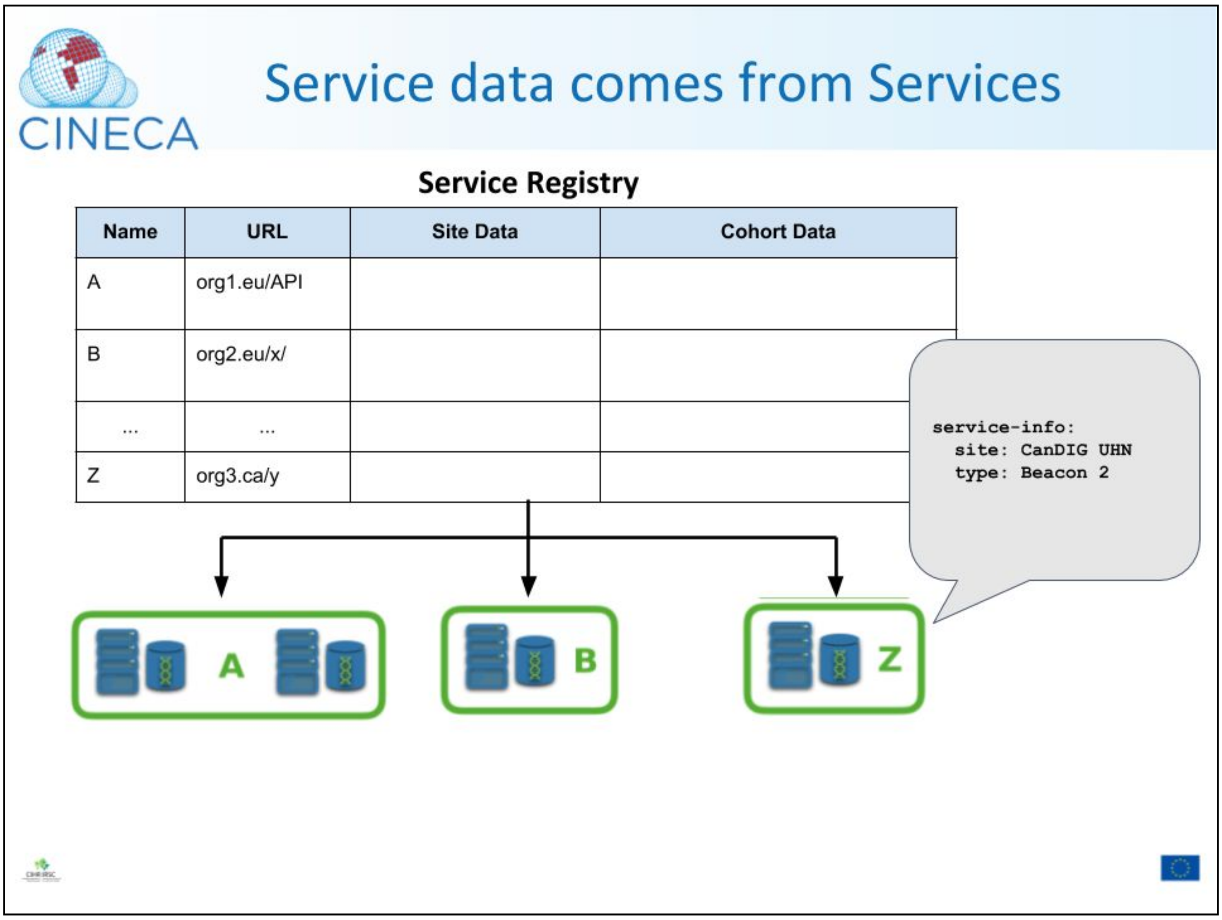
Service data comes from Services

Service Registry

Name	URL	Site Data	Cohort Data
A	org1.eu/API		
B	org2.eu/x/		
...	...		
Z	org3.ca/y		

Services are entered in the service registry with very basic information.

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



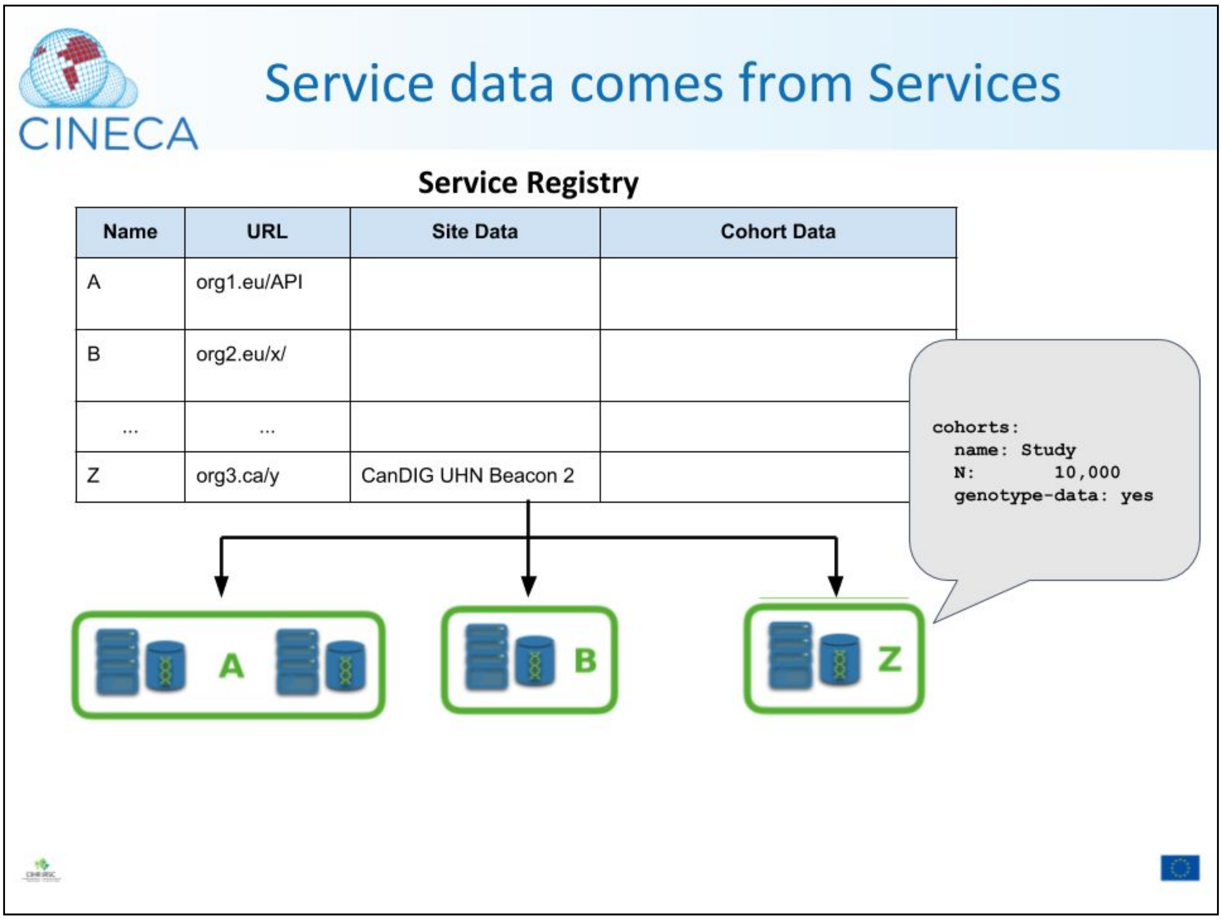
Service data comes from Services

Service Registry

Name	URL	Site Data	Cohort Data
A	org1.eu/API		
B	org2.eu/x/		
...	...		
Z	org3.ca/y		

```
service-info:
  site: CanDIG UHN
  type: Beacon 2
```

The services can be queried for their own information using the GA4GH service-info standard.



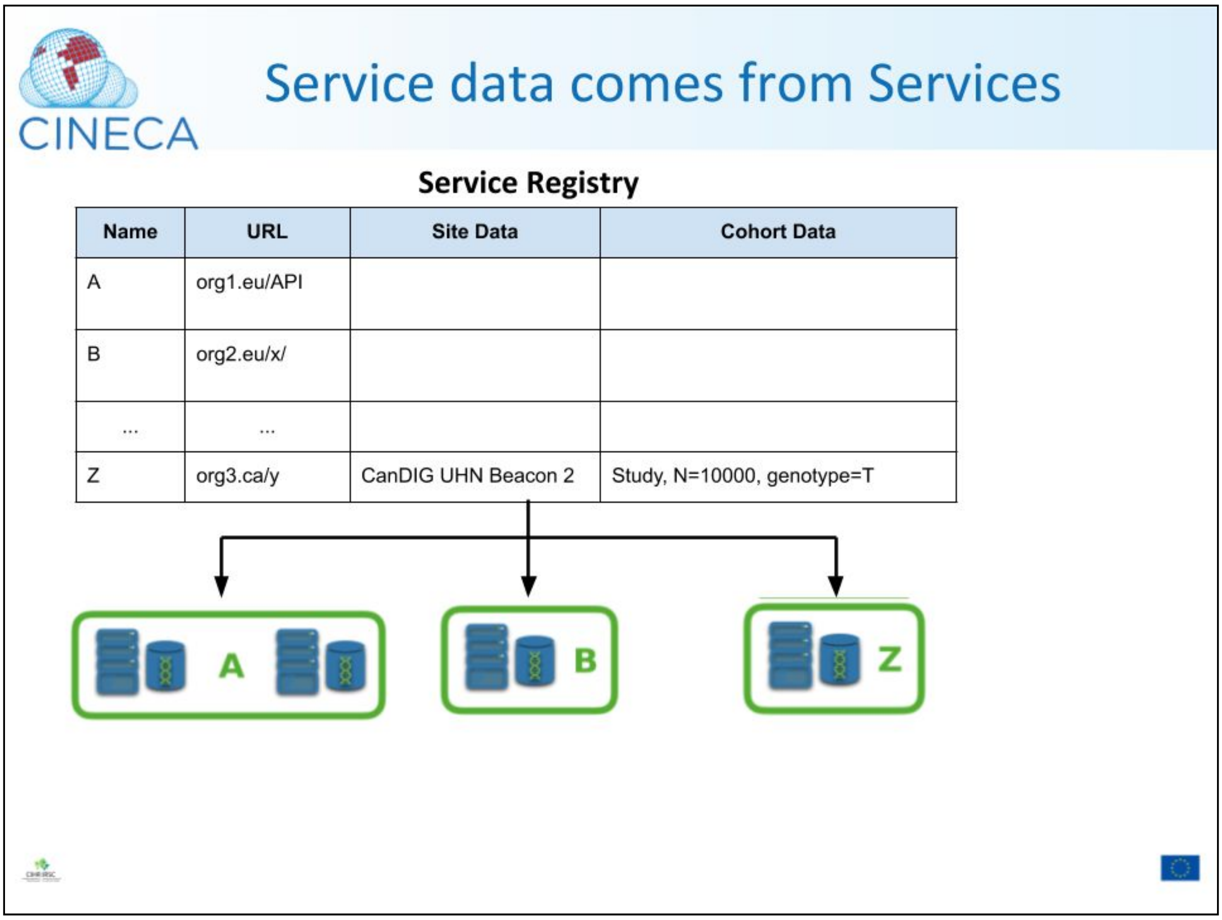
Service data comes from Services

Service Registry

Name	URL	Site Data	Cohort Data
A	org1.eu/API		
B	org2.eu/x/		
...	...		
Z	org3.ca/y	CanDIG UHN Beacon 2	

```
cohorts:
  name: Study
  N:        10,000
  genotype-data: yes
```

The services can also be queried for overall cohort information.



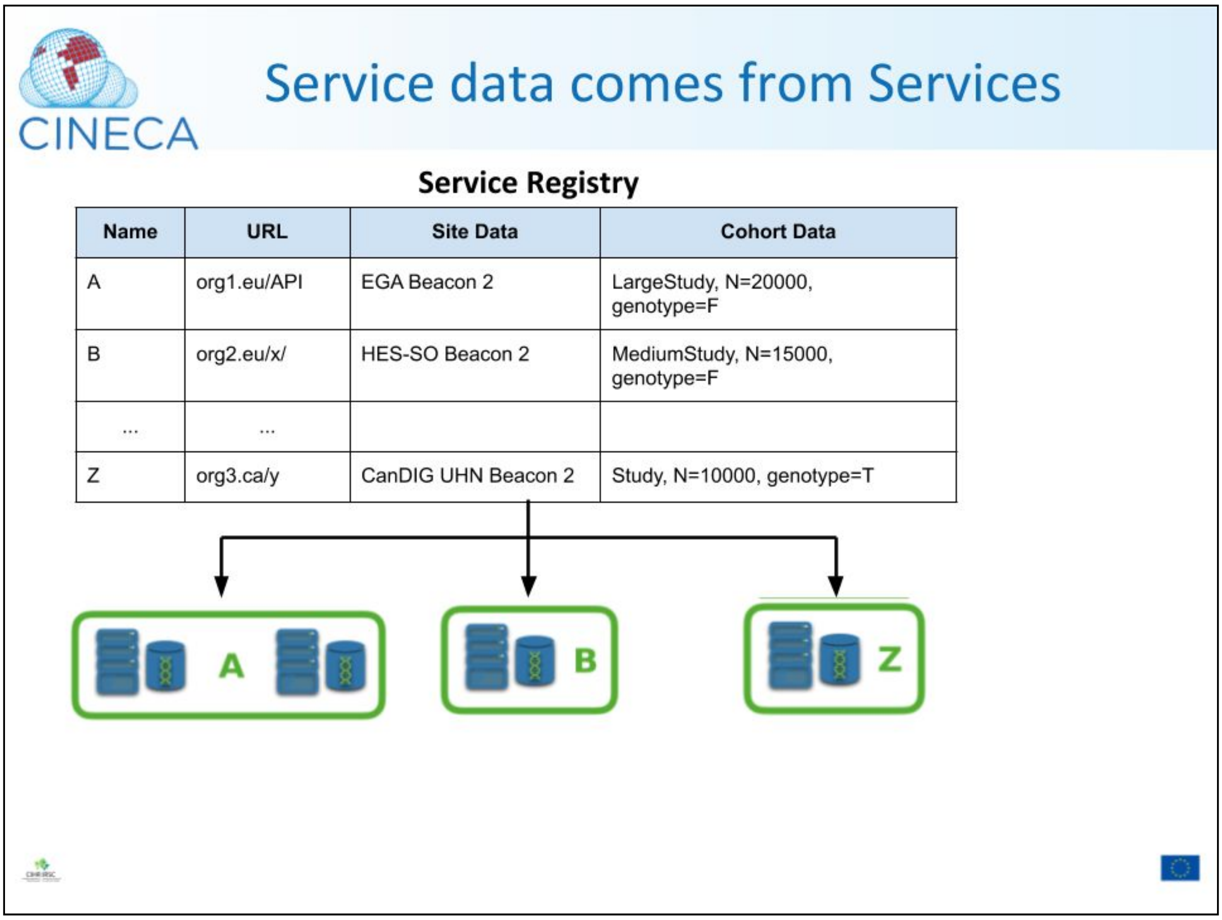
Service data comes from Services

Service Registry

Name	URL	Site Data	Cohort Data
A	org1.eu/API		
B	org2.eu/x/		
...	...		
Z	org3.ca/y	CanDIG UHN Beacon 2	Study, N=10000, genotype=T

That information is then added to the service registry.

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

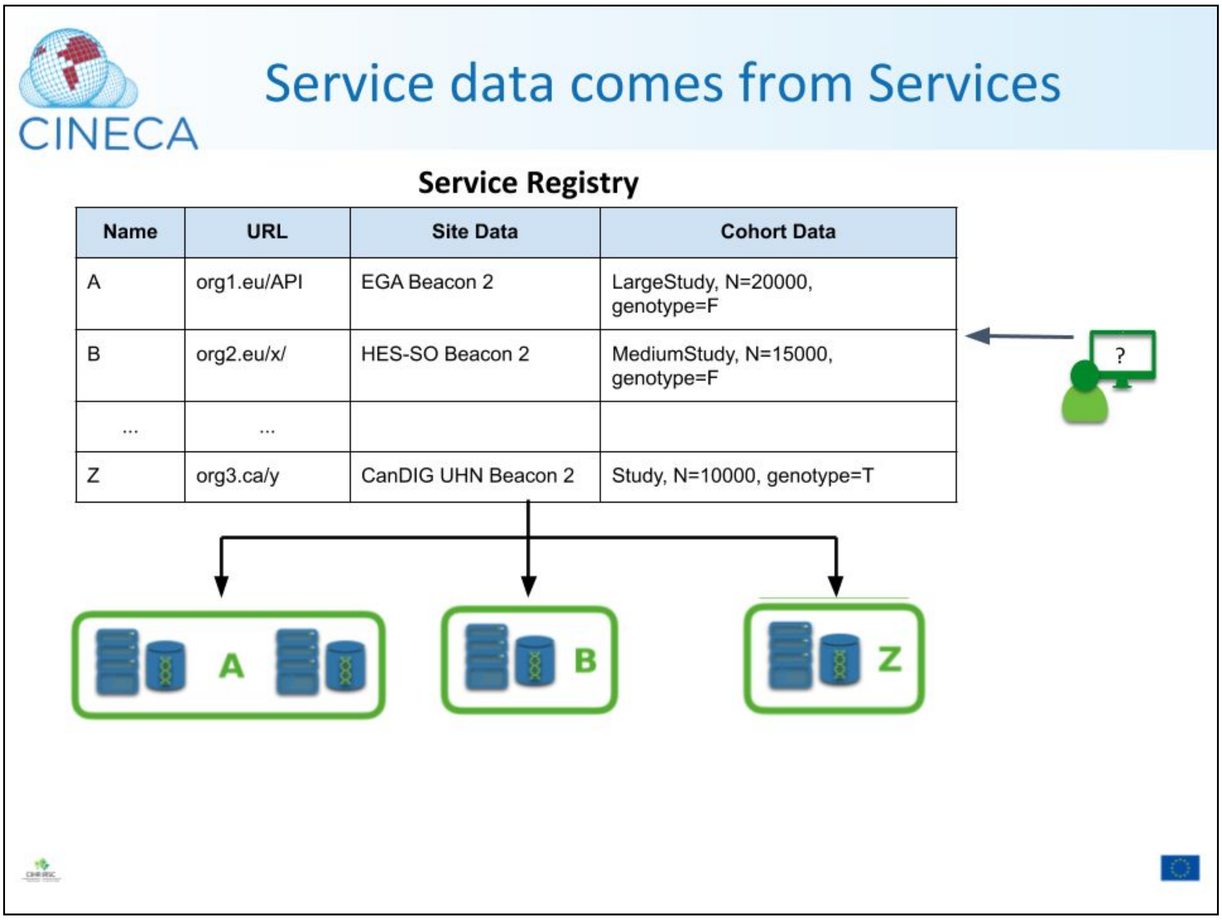


Service data comes from Services

Service Registry

Name	URL	Site Data	Cohort Data
A	org1.eu/API	EGA Beacon 2	LargeStudy, N=20000, genotype=F
B	org2.eu/x/	HES-SO Beacon 2	MediumStudy, N=15000, genotype=F
...	...		
Z	org3.ca/y	CanDIG UHN Beacon 2	Study, N=10000, genotype=T

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



Service data comes from Services

Service Registry

Name	URL	Site Data	Cohort Data
A	org1.eu/API	EGA Beacon 2	LargeStudy, N=20000, genotype=F
B	org2.eu/x/	HES-SO Beacon 2	MediumStudy, N=15000, genotype=F
...	...		
Z	org3.ca/y	CanDIG UHN Beacon 2	Study, N=10000, genotype=T

The service registry can be queried directly, and also used to pass through queries to the individual services.

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