



UNIVERSITÀ DEGLI STUDI
DI MILANO

ELIXIR, l'Infrastruttura di Ricerca europea per i dati delle scienze della vita.

Evento virtuale "European Open Science Cloud" - 15 Ottobre 2020

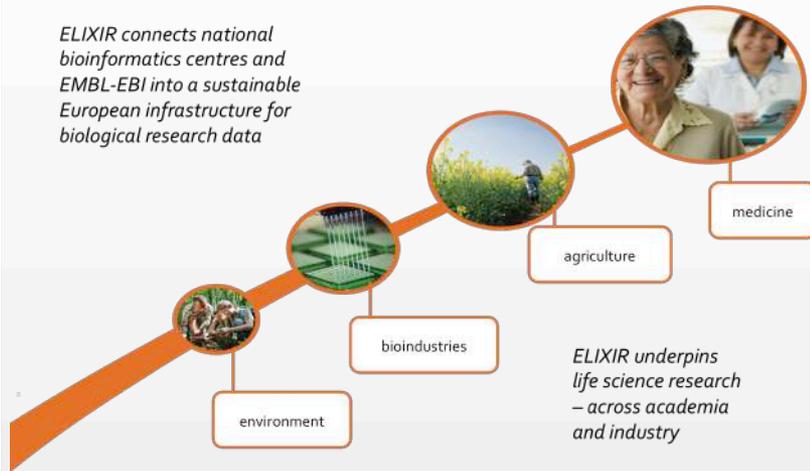


Federico Zambelli - ELIXIR-IT Technical Coordinator - RTD-B @



Cos'è ELIXIR?

ELIXIR connects national bioinformatics centres and EMBL-EBI into a sustainable European infrastructure for biological research data



- ELIXIR è l'infrastruttura Europea distribuita per i dati delle "Scienze della Vita".
- ELIXIR è una organizzazione intergovernativa che mette a fattor comune le risorse bioinformatiche in ambito Europeo, coordinandole ed integrandole.
- Per "risorse bioinformatiche" si intendono software (strumenti e librerie), banche dati, materiale per il training, buone pratiche, piattaforme computazionali e altro ancora.
- L'obiettivo di ELIXIR è far sì che queste risorse possano operare il più possibile come un'unica grande infrastruttura facilmente usufruibile da tutti i soggetti interessati: ricercatori, studenti, industrie, enti governativi, e semplici cittadini.

Perché serve? La sfida dei dati "omici"

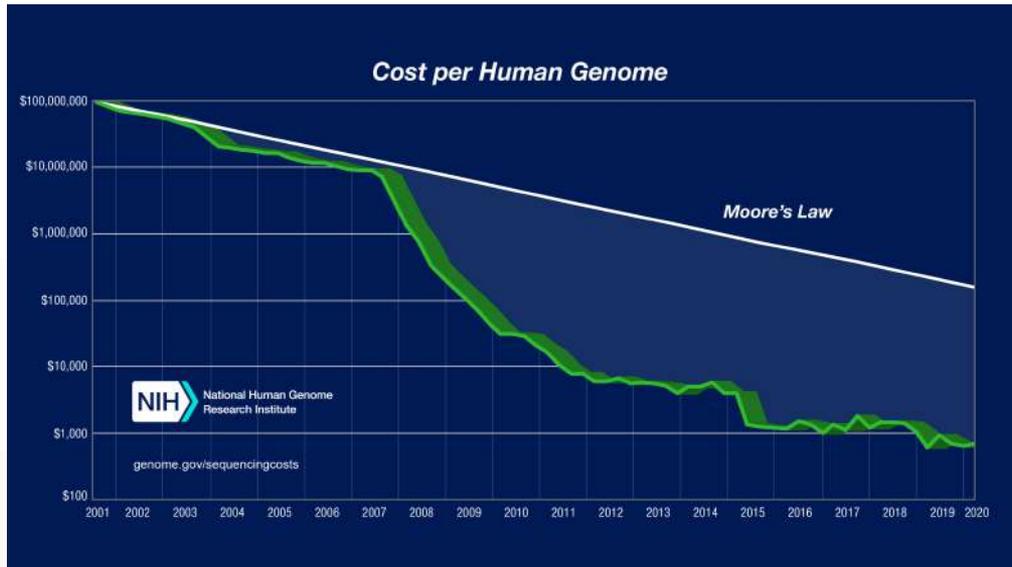


Immagine tratta da <https://www.genome.gov/sequencingcostsdata> (NIH)

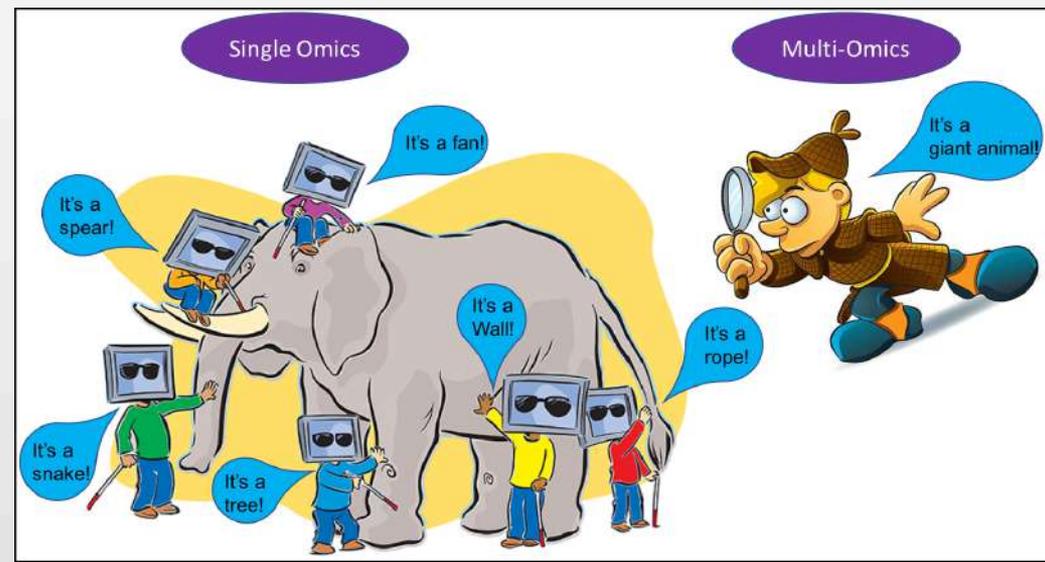
Lo sviluppo tecnologico ha trasformato profondamente lo studio degli aspetti molecolari della biologia, in particolare dalla seconda metà degli anni 2000 in poi, con l'avvento delle tecnologie di sequenziamento degli acidi nucleici dette NGS.

I biologi sono diventati rapidamente tra i più grandi produttori di dati scientifici in termini sia quantitativi...

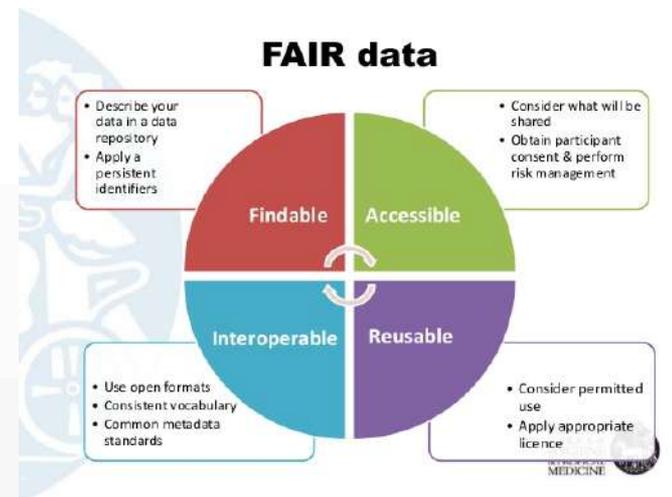
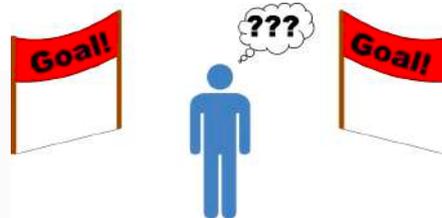
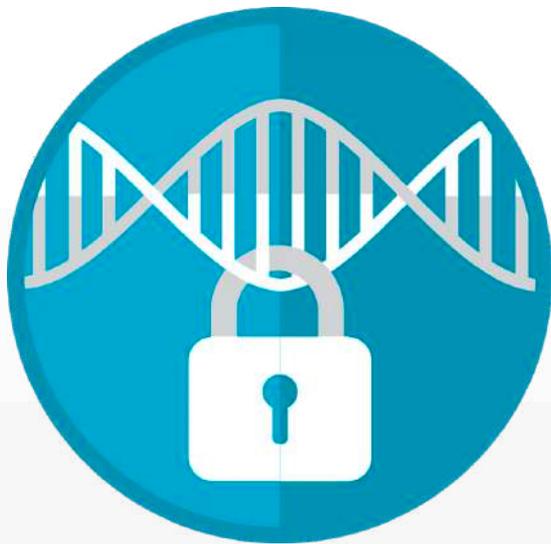
...che di complessità. Infatti le tecnologie hanno favorito lo sviluppo di approcci dapprima -omici e poi multi-omici.

Genomica, trascrittomica, epigenomica, proteomica, metabolomica, metagenomica, sono solo alcune delle innumerevoli scienze "omiche" che producono grandi quantità di dati.

Rendere aperti e fruibili questi dati significa agevolarne l'integrazione, e quindi l'estrazione delle preziose informazioni in essi contenute.



Perché serve? La sfida dei dati "omici"



Una sfida particolare differenzia poi alcune classi di dati biologici dalla maggior parte delle altre tipologie di dati scientifici.

I dati genomici umani (ed i metadati ad essi associati) sono dati estremamente sensibili e protetti dalle leggi di vari paesi e dal regolamento comunitario noto come GDPR.

Principio chiave: as open as possible, as closed as necessary.

Il modello organizzativo di ELIXIR



L'organizzazione di ELIXIR è di tipo distribuito e prevede un Hub di coordinamento presso Hinxton, UK e un Nodo per ogni nazione parte dell'organizzazione.

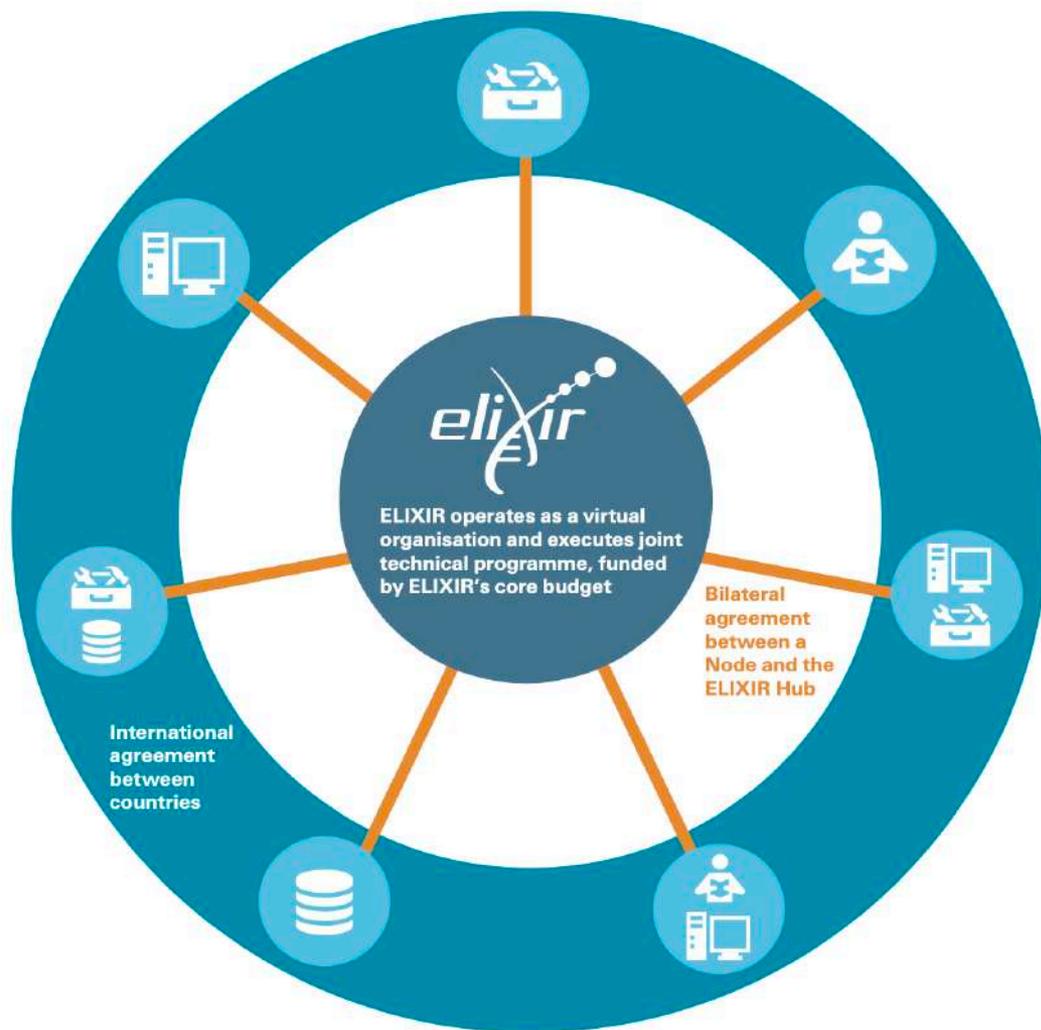
I Nodi nazionali, come quello italiano, ELIXIR-IT, decidono a loro volta come organizzarsi in base alla realtà locale.

Le nazioni più grandi hanno adottato a loro volta un modello distribuito per organizzare il nodo nazionale.

- 22 Members and 1 Observer
- Over 200 institutes
- Over 700 experts dedicated to infrastructure operations
- 18 Core Data Resources
- 145 services offered by ELIXIR Nodes
- 52 Implementation Studies
- Over 400 companies and SMEs engaged



Il modello organizzativo di ELIXIR

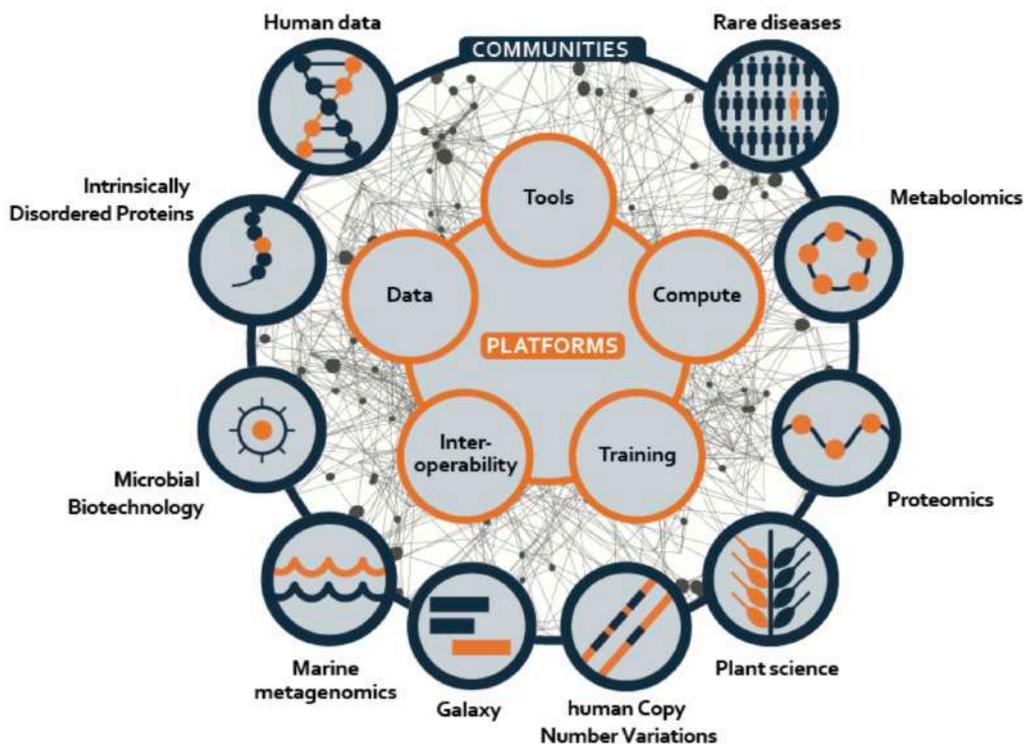


I Nodi nazionali mettono in comune i propri servizi nell'infrastruttura distribuita, ognuno secondo le competenze e le risorse disponibili localmente.

I servizi possono essere di diversi tipi, come centri di calcolo o altre risorse computazionali, banche dati o data repository, software, training.

Questo processo contribuisce alla crescita e al consolidamento delle infrastrutture nazionali dedicate ai dati biologici.

Il modello organizzativo di ELIXIR

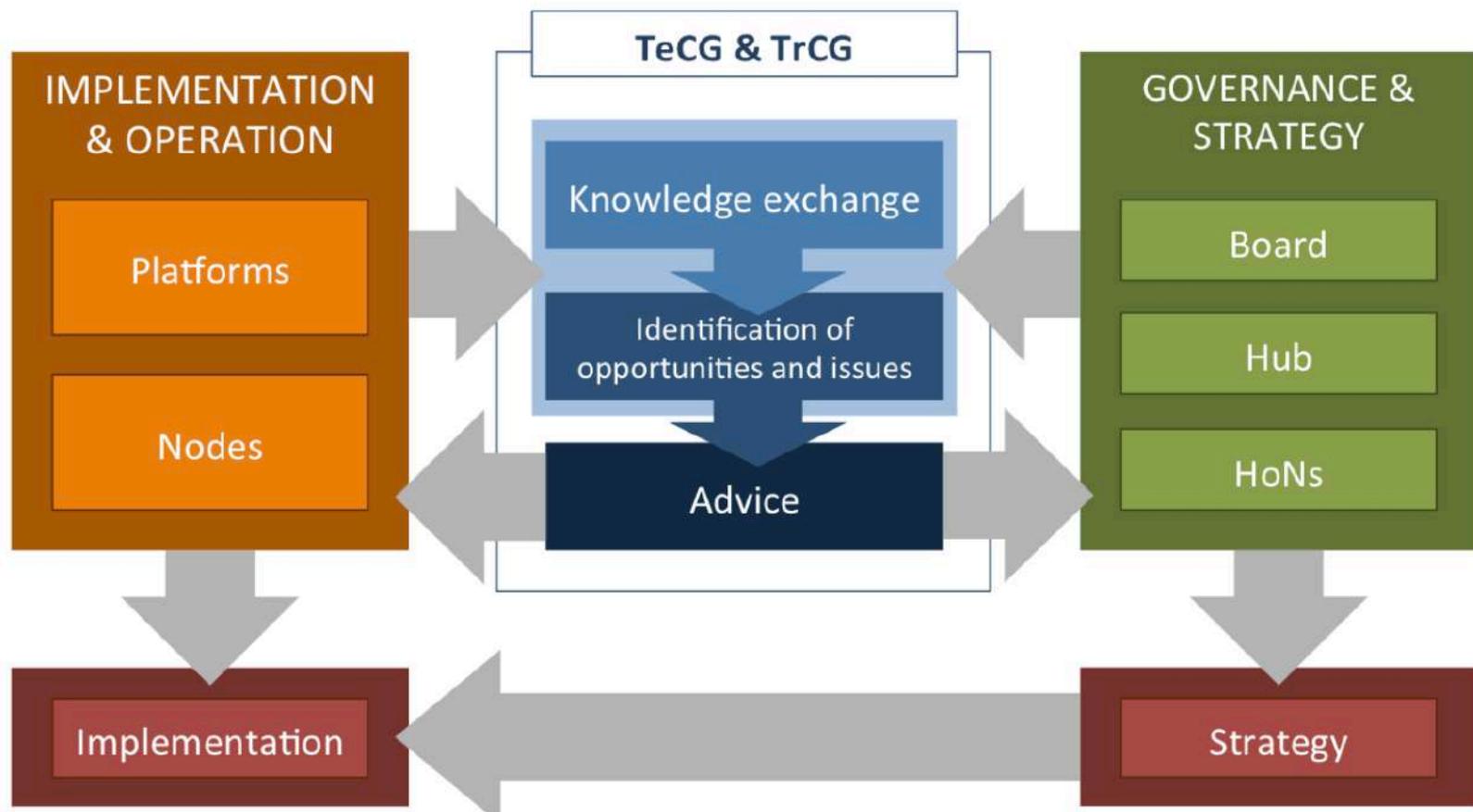


A loro volta, i servizi sono organizzati in cinque piattaforme che servono le comunità tematiche di ambito biologico e biomedico.

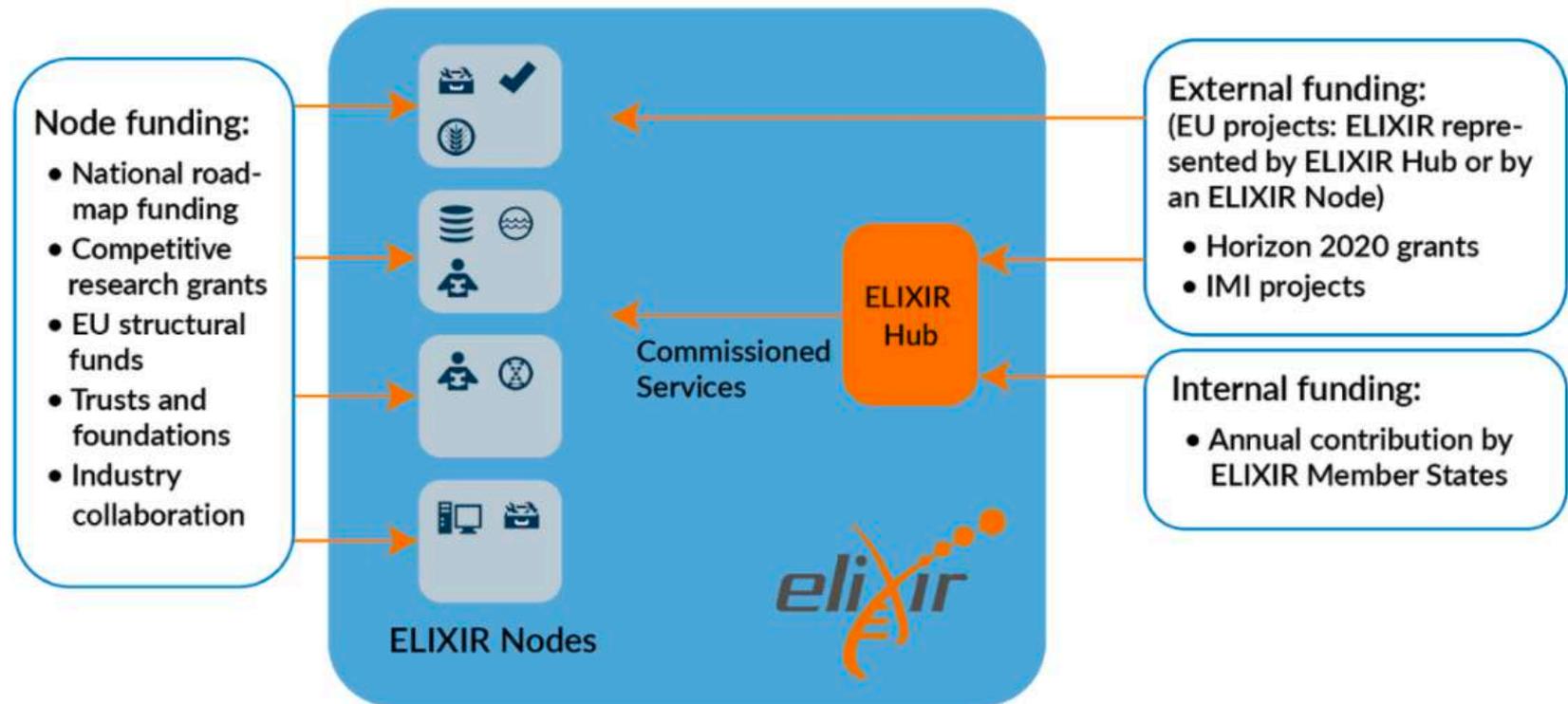
Le comunità sono entità "aperte" e possono essere partecipate da tutti i ricercatori che nutrano un interesse scientifico nelle varie tematiche.

Le comunità, di concerto con le piattaforme, possono proporre delle azioni volte a migliorare o far crescere l'infrastruttura attraverso degli "implementation study" a valere sul budget di ELIXIR.

Il modello organizzativo di ELIXIR



Come si sostiene ELIXIR?



I paesi membri pagano una quota annuale di partecipazione che copre le spese dell'Hub e permette in parte di finanziare gli implementation study/commissioned service. ELIXIR riceve fondi esterni partecipando a bandi di sviluppo infrastrutturale europei. I singoli nodi si finanziano attraverso national roadmap (non in Italia per ora) e anch'essi partecipando a bandi di sviluppo infrastrutturale nazionali o europei.

Alcuni risultati del quinquennio 2014-18

Platforms

 Data	 Tools	 Compute	 Interoperability	 Training
18 Core Data Resources identified	Over 10,000 tools registered in bio.tools	Demonstrated technology to transfer sensitive human data to secure cloud service	113 policies, 1128 databases, and 1199 standards metadata in FAIRsharing	Developed Training portal giving access to over 1000 training materials
12 ELIXIR recommended Deposition Databases	4,000 containers registered in BioContainers	ELIXIR AAI deployed with 67 services and over 1,800 registered users	Over 5,500,000 terms from 212 ontologies in Ontology Lookup Service	Developed Quality and impact assessment framework
Pioneered first external-review-based selection of Implementation Studies	Driving benchmarking within over 10 scientific communities	Drive development of European cloud capacities for life science (through EOSC)	655 collections, 822 resources with resolvable URIs and CURIEs prefixes in Identifiers.org	Organised 16 e-learning courses
	Drive development of Galaxy platforms		44 resources using Bioschemas metadata specifications	Organised 570 training events
				Trained 146 new bioinformatics trainers

Communities

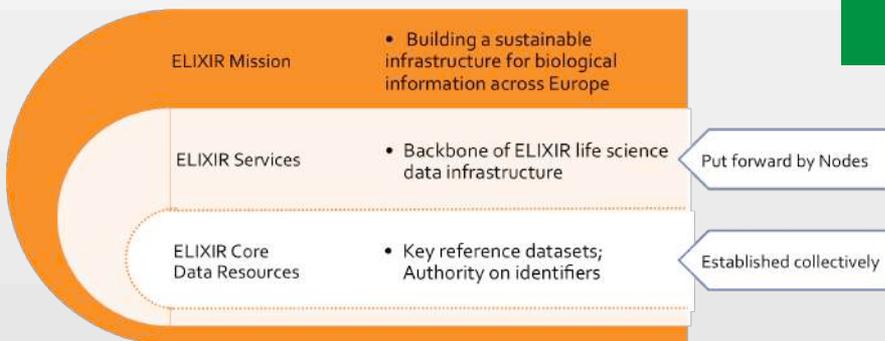
 Human data	 Rare diseases	 Marine	 Plants
Coordinating federated data management	Developed registry of over 100 data resources for rare disease research	Released ITSoneDB Marine section	Upgraded Minimum information about plant phenotyping experiments: MIAPPE
Developed ELIXIR Beacon implementation	Comparison and standardisation of services	Updated MGnifyDB: EBI Metagenomics Portal	Developed ELIXIR Plant Data lookup service
Increased sensitive data accessibility and re-use		Developed analysis pipelines for marine metagenomics (MGnifyDB and MMP)	Drive development of Breeding API & BrAPPS
		Released Ocean Gene Atlas	
		Launched Marine Metagenomics Portal (MMP)	

ELIXIR Core Data Resource

Le Core Data Resource sono banche dati biologiche identificate da ELIXIR come di **interesse strategico**.

L'identificazione di queste risorse avviene tramite una severa selezione che prende in considerazione diversi indicatori raggruppati in cinque categorie:

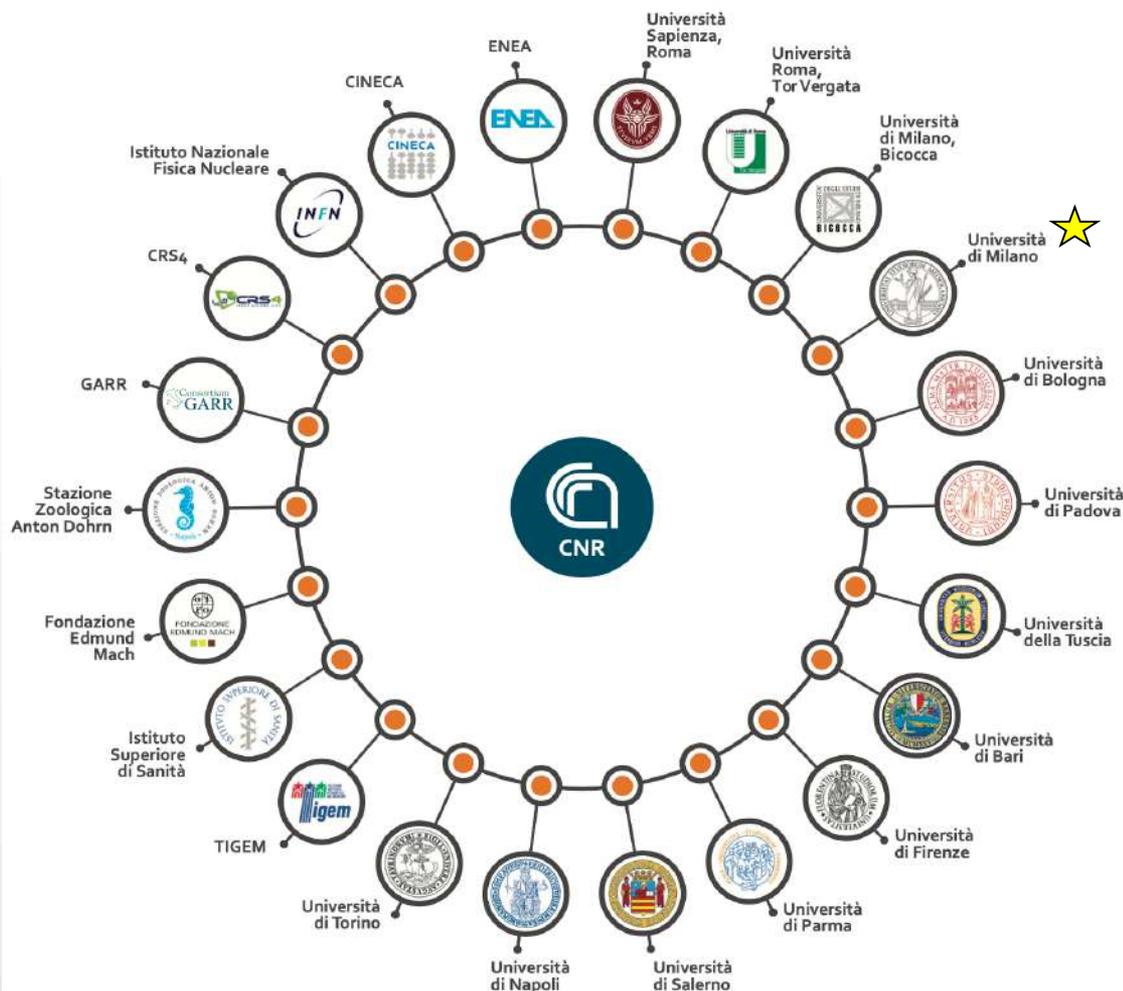
- Focus e qualità scientifica
- Rilevanza della comunità servita
- Qualità del servizio
- Governance e solidità a lungo termine
- Impatto



Core Data Resource	Data type
ArrayExpress	Functional Genomics Data from high-throughput functional genomics experiments.
BRENDA	Database of enzyme and enzyme-ligand information, across all taxonomic groups, manually extracted from primary literature and extended by text mining procedures, integration of external data and prediction algorithms.
CATH	A hierarchical domain classification of protein structures in the Protein Data Bank.
ChEBI	Dictionary of molecular entities focused on 'small' chemical compounds.
ChEMBL	Database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties and abstracted bioactivities.
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.
ENA	Nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
Ensembl	Genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation.
Ensembl Genomes	Comparative analysis, data mining and visualisation for the genomes of non-vertebrate species.
Europe PMC	Europe PMC is a repository, providing access to worldwide life sciences articles, books, patents and clinical guidelines.
Human Protein Atlas	The Human Protein Atlas contains information for a large majority of all human protein-coding genes regarding the expression and localization of the corresponding proteins based on both RNA and protein data.
The IMEx Consortium: represented by IntAct and MINT	IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. MINT focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators.
InterPro	Functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. <i>Note: This is an umbrella resource to which many collaborating databases contribute. In naming InterPro as a Core Data Resource, the critical role of the constituent databases is recognised.</i>
Orphadata	The Orphadata platform provides the scientific community with comprehensive, high-quality datasets related to rare diseases and orphan drugs, in a reusable and computable format.
PDBe	Biological macromolecular structures.
PRIDE	Mass spectrometry-based proteomics data, including peptide and protein expression information (identifications and quantification values) and the supporting mass spectra evidence.
SILVA	SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (Bacteria, Archaea and Eukarya).
STRING	Known and predicted protein-protein interactions.
UniProt	Comprehensive resource for protein sequence and annotation data.



ELIXIR-IT - JRU



Il Nodo italiano è configurato come una "Joint Research Unit" (JRU) che si prefigge di coordinare i servizi bioinformatici a livello nazionale e la loro integrazione in ELIXIR.

ELIXIR-IT diventa ufficialmente un nodo ELIXIR a gennaio 2016 dopo che il Ministero firma l'ELIXIR Consortium Agreement a dicembre 2015.

ELIXIR-IT è coordinato dal CNR e comprende un totale di 23 istituzioni tra Università, enti di ricerca e partner tecnologici.

ELIXIR-IT - un Nodo che cresce / membri

Original Node Application (2012)

Updated Node Application (2018)

Six members joined in 2019

Institution	Joined in
CNR (Lead)	2012
Sapienza Università di Roma	2012
Università di Roma Tor Vergata	2012
Università di Padova	2012
CINECA	2012
CRS4	2012
GARR	2012
Università di Milano 	2013
Università di Milano-Bicocca	2013
Università di Bologna	2013
Università della Tuscia	2013
INFN	2013
Università di Bari	2014
Università di Firenze	2016
Università di Parma	2016
Università di Salerno	2016
Fondazione Edmund Mach	2016
Università di Torino	2019
Università di Napoli Federico II	2019
Istituto Superiore di Sanità	2019
ENEA	2019
Tigem	2019
Stazione Zoologica Anton Dohrn	2019
Università di Verona	SAB cleared
Prospective New member (informal contacts)	2020?
...	...

ELIXIR-IT - un Nodo che cresce / servizi

Original Node Application
(2012)

6 Data Services

Compute Services from CINECA,
CRS4 and GARR

Training (generic commitment)

Updated Node Application
(2018)

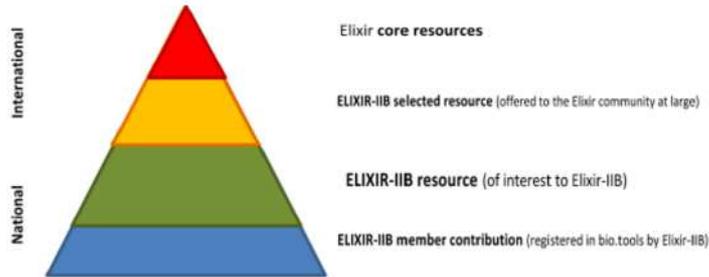
17 Data Services

26 Tools Services

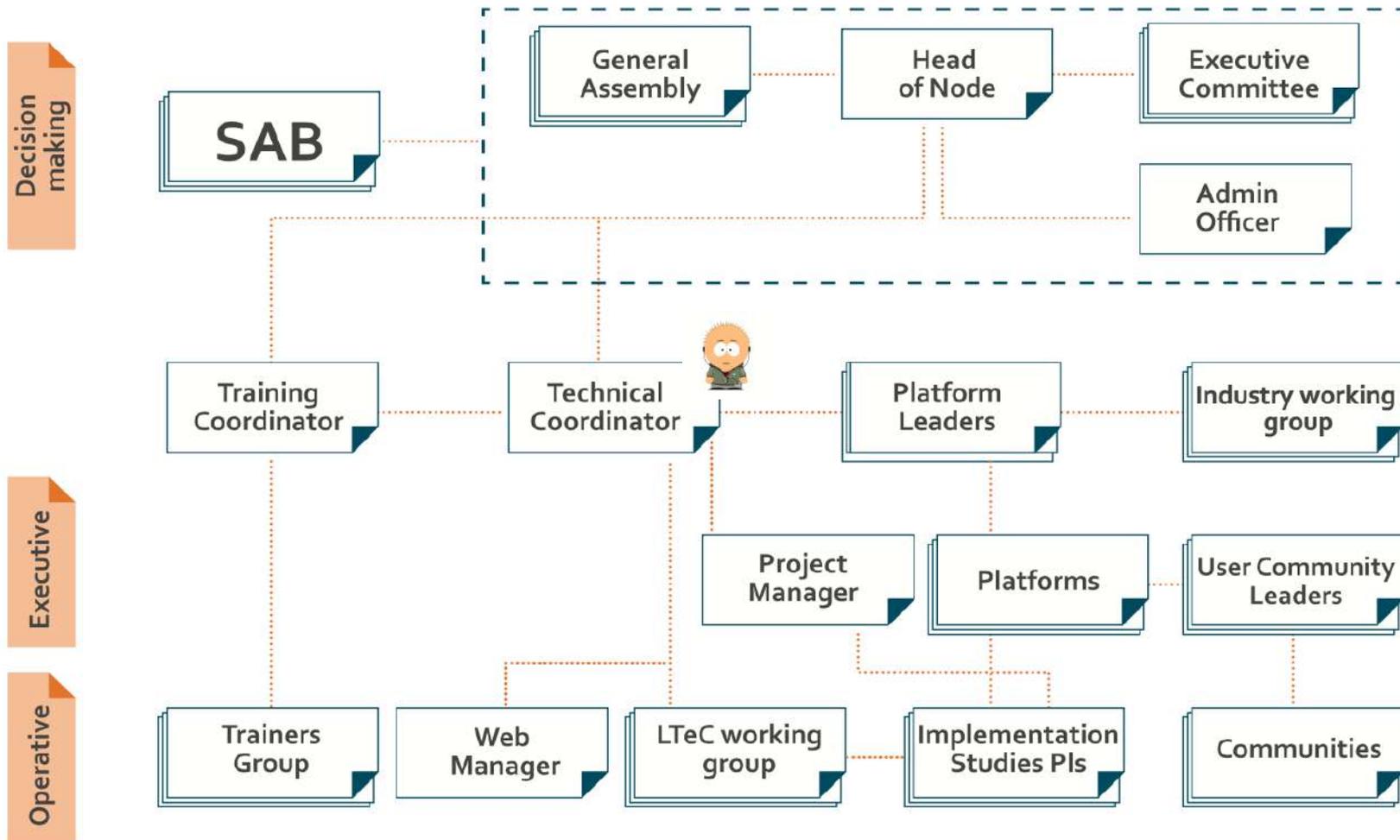
Improved Compute Services
from CINECA, INFN, GARR and
CRS4

3 Interoperability Services

Training: 21 Courses in 2017-18



ELIXIR-IT - Organizzazione interna



ELIXIR-IT - Es. Servizi 1 - HPC@Cineca

- Progetto pilota lanciato nel 2016
- ELIXIR-IT HPC@Cineca offre una fast-track di accesso alle risorse HPC del Cineca ai ricercatori europei.
- Gli utenti hanno a disposizione una piattaforma bioinformatica ricca di strumenti e comunque personalizzabile.
- L'accesso avviene tramite una call a sportello. I progetti sottomessi vengono valutati nel giro di 7gg lavorativi
- Il pacchetto base prevede 50K core hour e 5 Tb di spazio di archiviazione per una durata di sei mesi. Esigenze particolari vengono valutate caso per caso
- I progetti accedono ai più potenti cluster di calcolo disponibili ad oggi in Italia in ambito scientifico (Marconi A2 e Marconi 100).



Castrignanò et al. *BMC Bioinformatics* 2020, **21**(Suppl 10):352
<https://doi.org/10.1186/s12859-020-03565-8>

BMC Bioinformatics

SOFTWARE

Open Access

ELIXIR-IT HPC@CINECA: high performance computing resources for the bioinformatics community



Tiziana Castrignanò^{1*}, Silvia Gioiosa^{2,3}, Tiziano Flati^{2,3}, Mirko Cestari², Ernesto Picardi^{3,4}, Matteo Chiara^{3,5}, Maddalena Fratelli⁶, Stefano Amente⁷, Marco Cirilli⁸, Marco Antonio Tangaro⁹, Giovanni Chillemi^{3,9}, Graziano Pesole^{3,4*} and Federico Zambelli^{3,5*}



ELIXIR-IT - Es. Servizi 2 - Laniakea@ReCaS

elixir ITALY ELIXIR Italy Laniakea

HOME NEWS DOCUMENTATION USE ABOUT CONTACTS

LANIAKEA

Automatic deployment of virtual Galaxy environments for life science

LEARN MORE GET ACCESS USE

- Lanciato a febbraio di quest'anno
- Servizio di Galaxy on-demand. Galaxy è il più utilizzato workflow manager per la bioinformatica
- Rende l'analisi di dati bioinformatici facilmente fruibile e personalizzabile anche ad utenti non esperti
- Attualmente supporta progetti bioinformatici presso istituzioni quali l'Ospedale Gaslini di Genova, l'Istituto Ortopedico Rizzoli di Bologna, l'Istituto Zooprofilattico della Puglia e Basilicata, l' Estación Biológica de Doñana (Spagna) e altre.



TECHNICAL NOTE

Laniakea: an open solution to provide Galaxy “on-demand” instances over heterogeneous cloud infrastructures

Marco Antonio Tangaro ¹, Giacinto Donvito ², Marica Antonacci ², Matteo Chiara ³, Pietro Mandreoli ^{1,3}, Graziano Pesole ^{1,4} and Federico Zambelli ^{1,3,*}

ELIXIR e COVID19 / L'importanza dei dati Open

ELIXIR support to COVID-19 research

ELIXIR provides a range of services that you can use for studying the SARS-CoV-2 coronavirus and the COVID-19 disease. If you have questions about our COVID related services, please contact Kathi at katharina.lauer@elixir-europe.org.

- Find a database to store your data
- Access data relevant to COVID-19
- Make your data easier to find and share
- Find software and workflows to analyse your data
- Find computing resources to help you analyse datasets
- Contribute to ELIXIR's work on COVID-19
- Find COVID-19 publications from ELIXIR
- Services offered by other European infrastructures
- Find out more



Use our [interactive map](#) to find out what each country in ELIXIR is doing in response to COVID-19.

<https://elixir-europe.org/services/covid-19>



ELIXIR e COVID19 / L'importanza dei dati Open



[About](#) [Data Hubs](#) [Federated EGA](#) [Related resources](#) [Our partners](#) [Bulk downloads](#) [Submit data](#)

[Viral Sequences](#) [Host Sequences](#) [Expression](#) [Proteins](#) [Biochemistry](#) [Literature](#)

Host sequences

Accelerating research through data sharing



Viral sequences →

Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses.

[119,058 records >](#)

Host sequences →

Raw and assembled sequence and analysis of human and other hosts.

[997 records >](#)

About this portal

The COVID-19 Data Portal was launched in April 2020 to bring together relevant

The European COVID-19 Data Platform is developed and operated in a partnership of many institutions:

SARS-CoV-2 Data Hubs



Erasmus Medical Centre, the Netherlands

COVID-19 Data Portal



ELIXIR Europe

National Data Portals



Swedish COVID-19 Data Portal

<https://www.covid19dataportal.org/>



National Institute for Public Health and the Environment (RIVM), the Netherlands



EOSC-Life



ELIXIR-IT e COVID19 - CorGAT

Galaxy / CorGAT

Analyze Data Workflow Visualize Shared Data Help Login or Register Using 0 bytes

Tools

search tools

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

genome_alignment

Coronavirus Annotation Tool

Workflows

All workflows

Welcome, to CorGAT!

See the [CorGAT manual](#) for an explanation of what you can do with CorGAT

Configuring Galaxy » Installing Tools »

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

History

search datasets

Unnamed history (empty)

This history is empty. You can load your own data or get data from an external source

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

<http://corgat.cloud.ba.infn.it/galaxy>

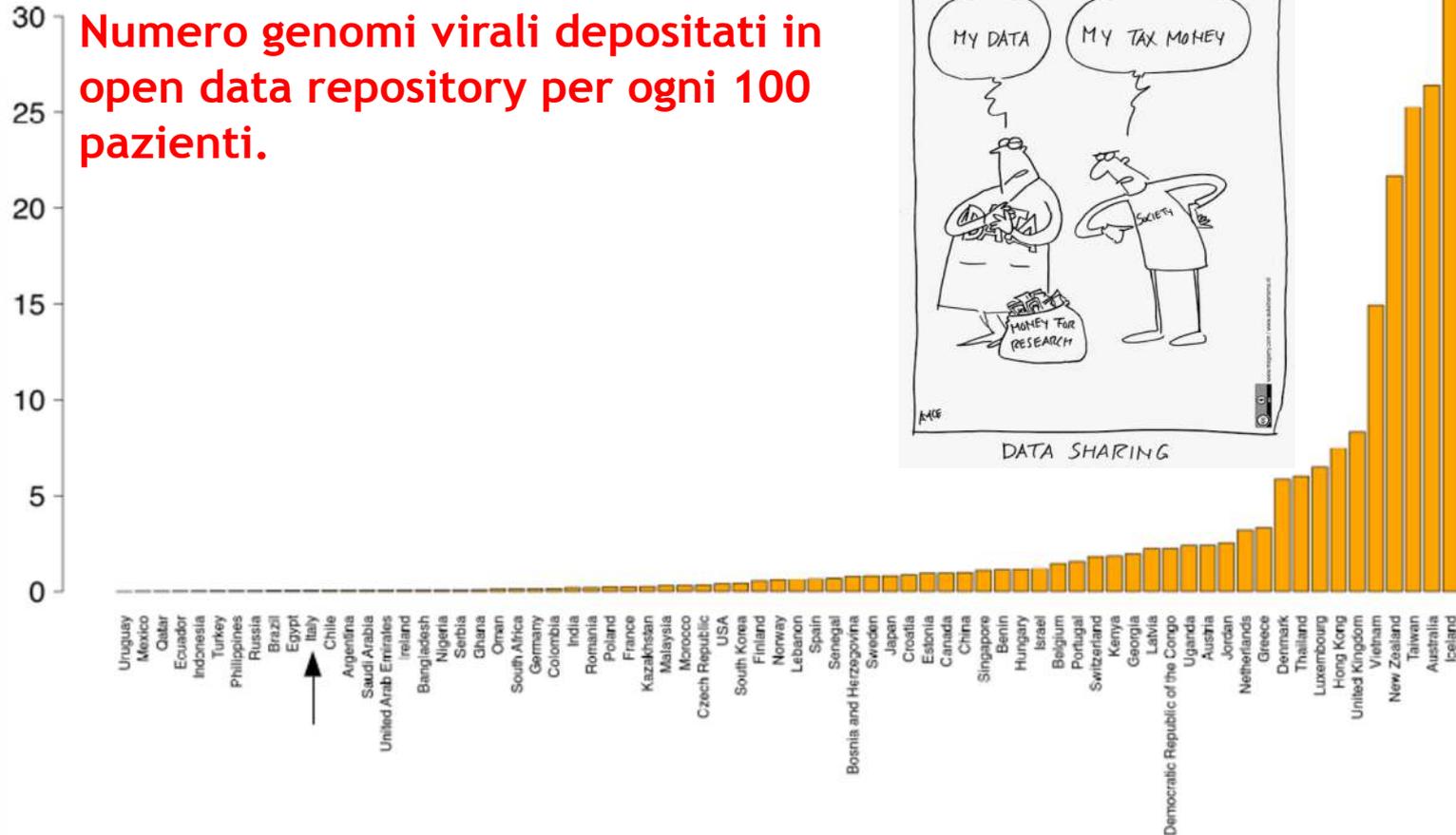
Strumento per l'annotazione funzionale dei genomi SARS-CoV-2



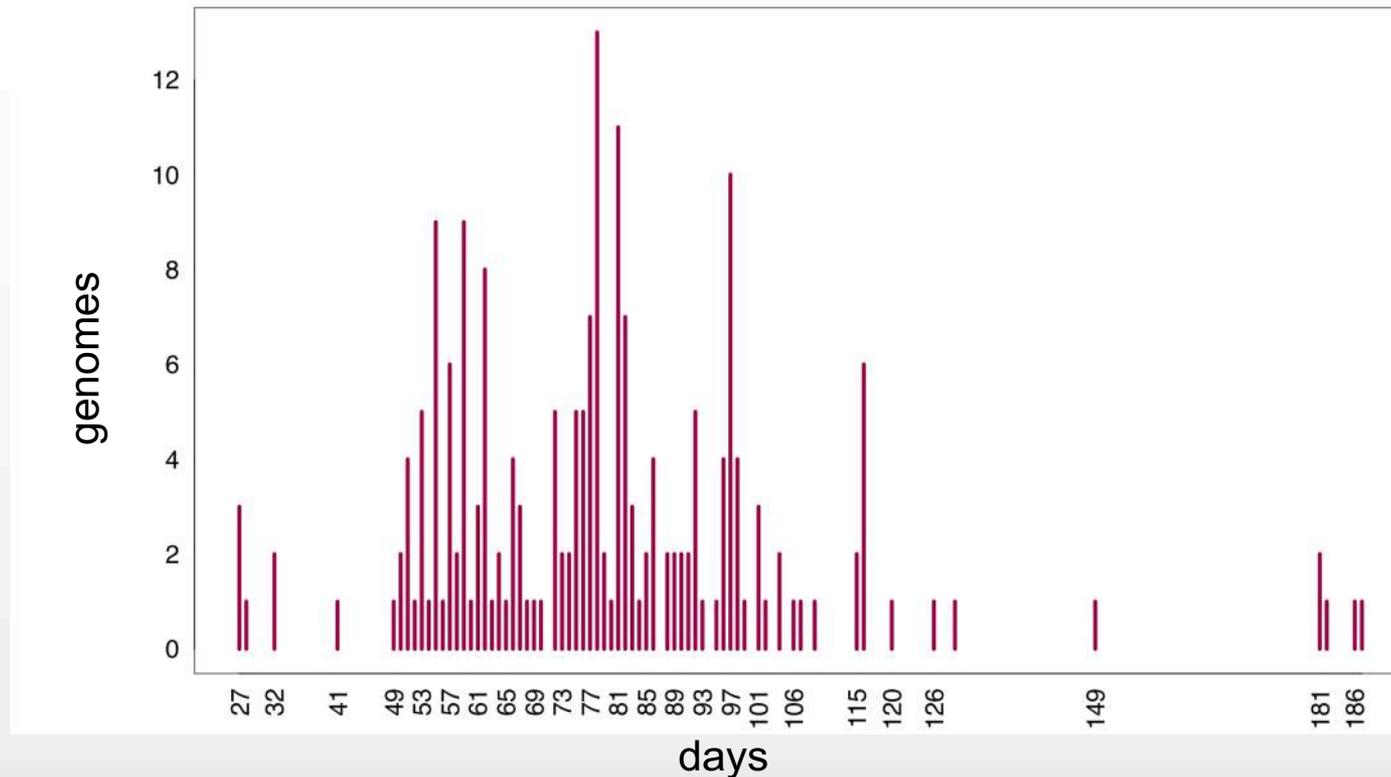
ELIXIR-IT e COVID19 - La battaglia per i dati

genomes per 100 cases

Numero genomi virali depositati in open data repository per ogni 100 pazienti.



ELIXIR-IT e COVID19 - La battaglia per i dati



Al momento è praticamente impossibile identificare l'emergere di un nuovo ceppo virale Sars-CoV-2 in Italia (soprattutto in tempo utile per adottare contromisure)

ELIXIR-IT @ Unimi - Partecipazione a Progetti Europei



Grazie!

<https://elixir-europe.org/>



ELIXIR-IT @ UNIMI

Prof. Giulio Pavesi giulio.pavesi@unimi.it

- Rappresentante UNIMI in Assemblea Generale
- Lead Piattaforma Tool ELIXIR-IT

Prof. David Horner david.horner@unimi.it

- Lead Human Data ELIXIR-IT

Dott. Matteo Chiara matteo.chiara@unimi.it

- Local Technical Coordinator UNIMI
- Co-Lead Human Data ELIXIR-IT

Dott. Federico Zambelli federico.zambelli@unimi.it

- Technical Coordinator
- Lead Galaxy Community ELIXIR-IT

Prof. Graziano Pesole
(UNIBA / CNR)

- Head of Node

Prof.ssa Rita Casadio
(UNIBO / CNR)

- Delegato MUR nel
ELIXIR Board
- Node Coordinator

Prof. Silvio Tosatto
(UNIPD)

- Deputy Head of Node
- Lead Piattaforma Data
ELIXIR

Dott.ssa Allegra Via
(CNR)

- Training Coordinator

Chi volesse maggiori informazioni o fosse interessato a partecipare alle attività del Nodo è benvenuto!

