1. Record Nr. UNINA9910770265603321

Autore	Bernasconi Anna
Titolo	Model, Integrate, Search... Repeat [[electronic resource]] : A Sound Approach to Building Integrated Repositories of Genomic Data / / by Anna Bernasconi
Pubbl/distr/stampa	Cham : , : Springer Nature Switzerland : , : Imprint : Springer, , 2023
ISBN	3-031-44907-X
Edizione	[1st ed. 2023.]
Descrizione fisica	1 online resource (277 pages)
Collana	Lecture Notes in Business Information Processing, , 1865-1356 ; ; 496
Disciplina	572.860285
Soggetti	Application software Bioinformatics Quantitative research Artificial intelligence - Data processing Computer and Information Systems Applications Data Analysis and Big Data Data Science
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di contenuto	Intro -- Foreword -- Preface -- Acknowledgements -- Contents -- Acronyms -- Introduction -- Genomic Data Integration -- Thesis Contribution within the GeCo Project -- The Recent COVID-19 Pandemic -- Thesis Structure -- Part I Human Genomic Data Integration -- Genomic Data Players and Processes -- Technological Pipeline of Genomic Data -- Production -- Integration -- Services and Access -- Taxonomy of Players Involved in Data Production and Integration -- Main Genomic Data Players -- Contributors -- Repository Hosts -- Consortia -- Integrators -- Modeling Genomic Data -- The Genomic Data Model -- The Genomic Conceptual Model -- Analysis of Metadata Attributes -- Model Design -- Validation: Source-Specific Views of GCM -- Related Works -- Integrating Genomic Data -- Theoretical Rationale -- Approach Overview -- Data Download -- Data Transformation -- Data Cleaning -- Data Mapping -- Data Normalization and Enrichment -- Search Service and Ontology Selection -- Enrichment Process -- Data Constraint Checking -- Architecture

Implementation -- Data Persistence -- Architecture Validation -- Lossless Integration -- Semantic Enrichment -- User Evaluation -- Related Works -- Snapshot of the Data Repository -- Included Data Sources -- OpenGDC: Integrating Cancer Genomic Data and Metadata -- Towards Automated Integration of Unstructured Metadata -- Experiments -- Related Works -- Searching Genomic Data -- Issues in Exploiting Semantic Knowledge in Genomics -- Inference Explanation -- GeKnowGraph: Exploration-Based Interface -- Exploration Interaction -- GenoSurf: a Web-Based Search Server for Genomic Datasets -- Data Search -- Key-Value Search -- Query Sessions -- Result Visualization -- Additional Functionalities -- Use Cases -- Validation of GenoSurf -- Related Works -- Future Directions of the Data Repository -- Including New Data Sources.
Improving Genomic Data Quality Dimensions -- Towards Better Interoperability -- Simplifying Data and Tools for End Users -- Monitoring Integration and Search Value -- Part II Viral Sequence Data Integration -- Viral Sequences Data Management Resources -- Landscape of Data Resources for Viral Sequences -- Fully Open-Source Resources -- GISAID and its Resources -- Integration of Sources of Viral Sequences -- Metadata Integration -- Value Harmonization and Ontological Efforts -- Replicated Sequences in Multiple Sources -- SARS-CoV-2 Search Systems -- Portals to NCBI and GISAID Resources -- Integrative Search Systems -- Comparison -- Discussion -- GISAID Restrictions -- Metadata Quality -- (Un)Willingness to Share Sequence Data -- Modeling Viral Sequence Data -- Conceptual Modeling for Viral Genomics -- Answering Complex Biological Queries -- Related Works -- Integrating Viral Sequence Data -- Database Content -- Relational Schema -- Data Import -- Annotation and Variant Calling -- Data Curation -- Searching Viral Sequence Data -- Requirements Analysis -- Lessons Learnt -- Web Interface -- Example Queries -- Discussion -- Related Works -- Future Directions of the Viral Sequence Repository -- Research Agenda -- ViruSurf Extensions -- Visualization Support: VirusViz -- Active Monitoring of SARS-CoV-2 Variations -- Integrating Host-Pathogen Information -- The Virus Genotype - Host Phenotype Connection -- The Host Genotype - Host Phenotype Connection -- Part III Epilogue -- Conclusions and Vision -- Summary of Thesis Contributions -- Achievements within GeCo Project -- Outlook -- META-BASE tool configuration -- User Manual -- Process Configuration -- Mapper Configuration -- Experimental setup GEO metadata extraction -- Mappings of viral sources attributes into ViruSurf -- References.

Sommario/riassunto

This book is a revised version of the PhD dissertation written by the author to receive her PhD from the Department of Electronics, Information and Bioengineering at Politecnico di Milano, Italy. The work deals with one of the central objectives of the European Research Council project “Data-Driven Genomic Computing”, i.e., building an integrated repository for genomic data. It reflects the research adventure that starts from modeling biological data, goes through the challenges of integrating complex data and their describing metadata and finally builds tools for searching the data empowered by a semantic layer. The results of this thesis are part of a broad vision: the availability of conceptual models, related databases, and search systems for both humans and viruses genomics will provide important opportunities for research, especially if virus data will be connected to its host, the human being, who is the provider of genomic and phenotype information. In 2023, the PhD dissertation won the CAiSE PhD Award, granted to outstanding PhD theses in the field of information systems engineering.