UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

DIPARTIMENTO DI SCIENZE MATEMATICHE, INFORMATICHE E FISICHE

TESI DI LAUREA IN
INFORMATICA

Modellazione e realizzazione di una base di dati per il monitoraggio del batterio legionella

CANDIDATO
Francesco Bortuzzo

RELATORE
Professor Angelo Montanari

CORRELATORE
Dottor Andrea Brunello
Dottor Nicola Saccomano

Anno Accademico 2023/2024

Indice

1. Introduzione e obiettivo del progetto

1.a. Introduzione al batterio Legionella

Il batterio Legionella è un bacillo gram-negativo aerobio, non mobile, che prospera in ambienti acquatici e umidi, sia naturali, come acque sorgive, termali, di fiumi o laghi, sia artificiali, come tubature, serbatoi, fontane e piscine. La Legionella è in grado di sopravvivere in una vasta gamma di condizioni ambientali, tra cui temperature comprese tra venti e quarantacinque gradi Celsius, pH neutro o leggermente alcalino, e presenza di nutrienti organici.

Il genere comprende sessantadue specie diverse, suddivise in settantuno sierotipi, di cui circa venti sono patogeni per l'uomo. La specie più comune è *Legionella pneumophila*, responsabile della maggior parte dei casi riportati di legionellosi[1]. La malattia può essere contratta inalando aerosol contenenti il batterio, come quelli prodotti da docce, fontane, impianti di condizionamento o umidificatori.

È quindi di fondamentale importanza monitorare la diffusione di questo batterio negli ambienti umidi e acquatici. Particolare attenzione deve essere rivolta alle strutture ospedaliere, termali e alberghiere, che rappresentano per loro natura ambienti a rischio di diffusione del batterio.

1.b. Legionella in Friuli Venezia Giulia

A livello europeo, la raccolta di dati relativi alla presenza del batterio è effettuata dall'ECDC[2]. Nel nostro Paese, invece, questa attività è svolta da diversi enti e istituzioni. Un contributo significativo proviene dall'Istituto Superiore di

[1] Legionellosi, o malattia del legionario, è una malattia infettiva che si presenta con sintomi simili all'influenza, come febbre, tosse, dolori muscolari e mal di testa. In alcuni casi, può evolvere in una forma polmonare, con sintomi analoghi a quelli della polmonite, e comportare complicazioni gravi, come polmonite atipica o decesso.

[2] Centro europeo per la prevenzione e il controllo delle malattie, istituito nel 2005.

Sanità e dai vari organismi che costituiscono il SNPA[3], di cui fa parte l'ARPA FVG[4].

I dati raccolti sono utilizzati per valutare il rischio di diffusione del batterio e adottare le misure di prevenzione e controllo indicate dal Ministero della Salute nelle "Linee guida per la prevenzione ed il controllo della legionellosi"[5].

In questo ambito, l'ARPA FVG ha effettuato numerose indagini sul territorio e ha pubblicato i risultati in vari report. Ad esempio, nel 2019, una collaborazione con l'Università degli Studi di Udine ha portato alla pubblicazione di un articolo[6], relativo alla presenza di Legionella nei sistemi di raccolta e distribuzione dell'acqua nella regione. Lo studio ha coperto un periodo di sedici anni, dal 2002 al 2017, durante il quale sono stati raccolti e analizzati 20.319 campioni attraverso 3.983 indagini ambientali.

I risultati riferiti alle indagini non clinciche e dunque eseguite routinariamente nell'ambito del piano regionale di sorveglianza ambientale hanno evidenziato che la presenza di Legionella è diffusa sopratutto nei cluster di impianti termali, nei quali il batterio è stato individuato nel 57,8% dei siti indagati, e in quelli ospedalieri, in cui nel 50,8% delle strutture è stata riscontrata la Legionella almeno una volta, con picchi dei campioni positivi soprattutto nei mesi che segnano l'inizio del periodo autunnale.

Inoltre, si è osservato che la presenza del batterio ha registrato un notevole incremento tra la seconda metà del 2006 e l'inizio del 2009, seguito da una diminuzione fino al 2013 e da un nuovo aumento negli anni successivi. Questo andamento indica chiaramente che, per ridurre il rischio di diffusione del batterio, è essenziale implementare un piano di prevenzione adeguato, che comprenda sia la manutenzione degli impianti sia la sorveglianza ambientale.

1.c. Obiettivo del progetto

Pur riconoscendo l'importanza cruciale della sorveglianza ambientale per il controllo della legionellosi, in Friuli Venezia Giulia, come in molte altre regioni, manca un sistema efficiente per la memorizzazione, la gestione e l'analisi dei dati raccolti. Tale carenza rende estremamente oneroso lavorare con la mole di

[3]Sistema Nazionale per la Protezione dell'Ambiente

[4]Agenzia Regionale per la Protezione dell'Ambiente Friuli Venezia Giulia.

[5][1] M. della Salute, «Linee guida per la prevenzione e il controllo della legionellosi», 2015, [Online]. Disponibile su: https://www.salute.gov.it/portale/malattieInfettive/dettaglioPubblicazioniMalattieInfettive.jsp?id=2362

[6][2] A. Felice, M. Franchi, S. De Martin, N. Vitacolonna, L. Iacumin, e M. Civilini, «Environmental surveillance and spatio-temporal analysis of Legionella spp. in a region of northeastern Italy (2002–2017)», *PLOS ONE*, vol. 14, fasc. 7, p. e218687, 2019, doi: 10.1371/journal.pone.0218687.

informazioni raccolte nelle indagini ambientali, ostacolando così lo svolgimento di analisi e ricerche mirate.

In questo contesto, i sistemi di basi di dati giocano un ruolo fondamentale, in quanto permettono di memorizzare grandi quantità di dati e di effettuare ricerche complesse in modo rapido ed efficiente.

Il presente documento si propone di delineare gli aspetti principali per la progettazione di un database relazionale destinato alla memorizzazione dei dati relativi alla diffusione della Legionella. Più specificamente, nei prossimi capitoli viene condotta un'analisi critica di una soluzione esistente, rispetto alla quale sono proposte alcune modifiche al fine di adattarla alle nuove esigenze emerse dai colloqui condotti in collaborazione con i ricercatori dell'ARPA FVG. Successivamente, vengono descritte le fasi di ristrutturazione, traduzione in modello relazionale e implementazione della base di dati, con particolare attenzione rivolta alla definizione dei domini e dei vincoli che garantiscono l'integrità dei dati.

2. Analisi critica di una soluzione pre-esistente: Valutazione e proposte di modifica

Come accennato nel capitolo introduttivo, una delle principali sfide riscontrate nell'attuale sistema di gestione dei dati riguarda la realizzazione di soluzioni efficienti per la memorizzazione delle informazioni raccolte durante le indagini ambientali. In questa sezione si procede a un'analisi critica di un database relazionale utilizzato per archiviare i dati relativi alla diffusione della Legionella. Il database oggetto di analisi è stato sviluppato dal dottor Dario Garlatti nell'ambito della sua tesi di laurea triennale in informatica, dal titolo "Base di dati e applicazione web per il monitoraggio del batterio della Legionella"[7].

2.a. Requisiti

Prima di procedere con lo studio del database, è necessario definire i requisiti del sistema informativo. Questi sono di natura qualitativa e descrivono le caratteristiche che il sistema deve possedere per soddisfare le esigenze degli utenti e degli stakeholder. I criteri alla base della progettazione della soluzione in analisi riguardano l'intera fase di acquisizione dei dati relativi alle indagini ambientali portate a termine dai ricercatori di ARPA FVG per il monitoraggio della Legionella in regione.

Di seguito sono riportati i requisiti, non strutturati, che hanno guidato la progettazione della base di dati.

Il sistema deve consentire la registrazione delle indagini ambientali relative alla presenza di Legionella nei sistemi di adduzione e conservazione dell'acqua. Ogni indagine è definita dal tipo, dalla data e dal sito presso il quale viene eseguita, ed è, eventualmente, associata al richiedente qualora si tratti di un'indagine di follow-up. Un sito è identificato dall'indirizzo e dalla categoria di appartenenza. Le indagini comprendono il prelievo di campioni, ciascuno dei quali è associato ad una e una sola indagine. Tali campioni sono caratterizzati dalla

[7] [3] D. Garlatti, «Base di dati e applicazione web per il monitoraggio del batterio della legionella», 2020.

temperatura, caldo o freddo, al momento dell'estrazione e dal punto di prelievo, all'interno del sito presso cui è svolta l'indagine cui afferiscono, e sono univocamente identificati da un codice. Tutti i campioni prelevati sono sottoposti a diverse analisi per accertare la presenza di Legionella. Si annoverano: la PCR qualitativa, che consente di rilevare la presenza del DNA del batterio; la PCR quantitativa, che misura la concentrazione di Legionella nei campioni, espressa in µg/l; l'analisi colturale, che consente di isolare e identificare le unità formanti colonia (UFC_L) e, in caso di positività, di determinare il sierogruppo.

2.A.I. Note

Si segnala che la PCR non costituisce un metodo diagnostico definitivo per la legionellosi, ma piuttosto un test di screening che necessita di conferma attraverso la coltura. Infatti, «poiché, così come specificato nella norma ISO "*Water quality- Detection and quantification of Legionella spp and/or Legionella pneumophila by concentration and genic amplification by quantitative polymerase chain reaction (qPCR)*" (ISO/TS 12869, 2012), la qPCR non da informazione riguardo lo stato delle cellule, la quantificazione dovrà sempre essere determinata mediante esame colturale»[8].

Inoltre, si osserva che i metodi analitici utilizzati per la rilevazione del batterio, come indicato nell'allegato 4 delle "Linee Guida per la prevenzione e il controllo della legionellosi"[9], variano in base alla matrice da analizzare (acqua, biofilm, aria); tuttavia, i risultati ottenuti sono espressi in modo uniforme, a prescindere dal tipo di analisi effettuata. Pertanto, considerata l'esigenza di conservare le informazioni relative ai risultati delle analisi sui campioni, si ritiene lecito mantenere le tre tipologie di analisi sopra menzionate, senza ulteriori distinzioni.

2.B. Schema concettuale-logico

Di seguito viene presentato lo schema concettuale-logico del database sviluppato dal dottor Garlatti. Tale schema è stato modellato utilizzando il linguaggio IDEF1X[10]. Questo linguaggio appartiene alla famiglia dei linguaggi di modellazione IDEF[11]. Per una corretta comprensione dello schema, è essenziale definire i concetti di entità e relazione, che rappresentano i fondamenti della modellazione dei dati.

[8] [1],«Linee guida per la prevenzione ed il controllo della legionellosi», p. 21
[9] [1], «Linee guida per la prevenzione ed il controllo della legionellosi», p. 91
[10] Integration DEFinition for information modeling.
[11] https://www.idef.com/

2.B.I. Notazione IDEF1X

Nella notazione IDEF1X, le entità sono rappresentate attraverso tabelle contenenti attributi che ne descrivono le proprietà, e ciascuna entità è identificata da una chiave primaria, costituita da un singolo attributo o da una combinazione di attributi in grado di identificare univocamente ogni riga della tabella. Un'entità può essere classificata come indipendente se può essere identificata senza necessità di relazioni con altre entità, mentre si considera dipendente quando il suo significato emerge solo in relazione a un'altra tabella associata.

Le relazioni di connessione, o associazioni, sono rappresentate mediante linee che collegano due entità, segnalando l'esistenza di un legame tra di esse. In particolare, si distinguono due tipi di relazioni: le associazioni identificative, in cui l'entità figlia è identificata in relazione all'entità genitore e la cui chiave primaria include quella del genitore, rappresentate da una linea continua; le associazioni non identificative, in cui l'entità figlia è comunque identificata in relazione all'entità genitore, ma la chiave primaria della figlia non include quella del genitore, rappresentate da una linea tratteggiata. La cardinalità di queste associazioni è indicata da lettere: "p" denota una relazione uno a uno o uno a molti, "z" indica una relazione uno a zero o uno a uno, e "n" specifica una relazione uno a esattamente n.

Le relazioni di categorizzazione, invece, sono rappresentate da linee che collegano un'entità genitore a una o più entità figlie, sottolineando che queste ultime ereditano le proprietà dell'entità genitore, pur mantenendo attributi distintivi. Le entità di categoria[12] sono mutuamente esclusive e si distinguono grazie a un attributo discriminatore, il cui valore è univoco per ciascuna entità di categoria. Esistono due tipologie di categorizzazione: le categorizzazioni complete, in cui ogni entità genitore deve essere associata a una figlia, rappresentate da un pallino vuoto e due linee; le categorizzazioni incomplete, in cui un'entità genitore può non essere associata a nessuna entità figlia, rappresentate da un pallino pieno e una linea.

[12]Entità che costituisce un sottotipo di un'altra.

2.B.II. Schema concettuale-logico

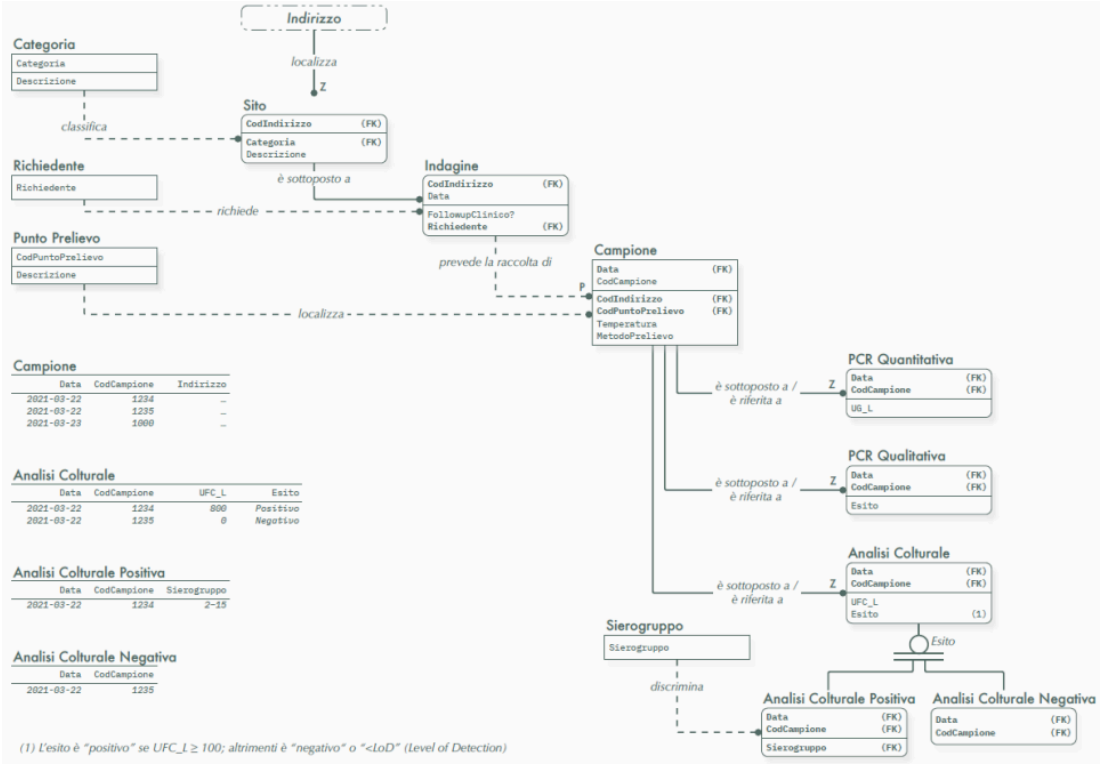

Figura 1: Diagramma ER

2.C. Analisi dello schema: Considerazioni e proposte di modifica

Lo schema illustrato è stato concepito per rispondere ai requisiti di memorizzazione dei dati relativi alla diffusione della Legionella. Tuttavia, durante una prima fase di analisi del database, sono stati individuati alcuni difetti che richiedono un'accurata valutazione e una eventuale revisione dello schema.

Alcune entità, come *indirizzo* e *categoria*, sono state inizialmente progettate come entità autonome, ma potrebbe essere più efficace trattarle come attributi dell'entità *sito*. Questo approccio non solo semplificherebbe lo schema, ma migliorerebbe anche la sua chiarezza strutturale. In particolare, l'attributo descrizione dell'entità *categoria* è superfluo, poiché il nome della categoria dovrebbe essere sufficiente per identificarla in modo univoco. Inoltre, l'aggiunta di un attributo nome all'entità *sito* potrebbe facilitare la consultazione dei dati, specialmente per quanto riguarda gli ospedali, che sono generalmente riconosciuti dalla combinazione di nome e città, piuttosto che unicamente dall'indirizzo. In aggiunta, si propone di arricchire l'entità sito con nuovi attributi che ne descrivano le caratteristiche principali nel contesto specifico. Questi attributi

includono dettagli sull'impiantistica del sito, come la tipologia di caldaia, il materiale delle tubature, l'uso del cloro, e altre informazioni di carattere generale, come l'anno dell'ultima ristrutturazione.

Un ulteriore elemento di riflessione riguarda l'associazione del *richiedente* alle *indagini ambientali*. Superando quanto indicato nei requisiti, si ritiene opportuno che l'entità *richiedente* sia messa in relazione con indagini che non siano unicamente di follow-up. Inoltre, si suggerisce l'introduzione di una nuova entità denominata *follow-up clinico*, associata a una o più indagini ambientali. Questa modifica si dimostra particolarmente efficace nella gestione dei dati relativi ai pazienti affetti da legionellosi e nella valutazione del rischio di diffusione del batterio. Infatti, «per avere un quadro globale della situazione, è fondamentale disporre, per ciascun paziente affetto da legionellosi, di informazioni precise su una eventuale esposizione a rischio nei dieci giorni precedenti l'insorgenza dei sintomi»[13]. La possibilità di associare un paziente a una o più indagini ambientali risulterebbe, dunque, vantaggiosa.

L'entità *follow-up clinico* potrebbe essere ulteriormente arricchita con attributi volti a descrivere il paziente e la sua esposizione al rischio, quali la data di insorgenza dei sintomi, il luogo di residenza, il luogo di lavoro e le attività svolte nei dieci giorni precedenti l'insorgenza dei sintomi. Questi dettagli, tuttavia, non sono modellati nello schema attuale né saranno inclusi nello schema finale, poiché non sono stati considerati nei requisiti né approfonditi con i ricercatori. Ciononostante, potrebbero rivelarsi utili per una valutazione più accurata del rischio di diffusione del batterio.

Per quanto concerne l'entità *campione*, è opportuno valutare l'introduzione di un attributo volume per specificare la quantità d'acqua prelevata per l'analisi. Sebbene non strettamente necessario, tale attributo trova pertinenza nel definire parametri di riferimento relativi al prelievo dei campioni, come il volume minimo d'acqua richiesto per eseguire tutte le analisi previste. Inoltre, poiché è possibile prelevare campioni di diversa matrice ambientale, come acqua, biofilm o aria, si presenta la proposta di introdurre un attributo "matrice" che consenta di specificare il tipo di campione analizzato.

Infine, si propone di riorganizzare la disposizione delle entità *indagine ambientale* e *campione* all'interno dello schema. In particolare, per come definita nella sezione 8, un'indagine ambientale non è altro che una collezione di campioni prelevati in un sito specifico in una data determinata. Pertanto, risulta più coerente associare solo l'entità *campione* alle informazioni spaziali contenute nelle tabelle *punto di prelievo* e *sito*. Si noti che tale modifica comporta l'introduzione di un vincolo di integrità che stabilisce che tutti i campioni associati a un'indagine devono essere prelevati nello stesso sito.

[13][1], «Linee guida per la prevenzione ed il controllo della legionellosi», p. 30

In questo contesto, appare vantaggioso apportare una modifica alla struttura delle entità *sito* e *punto di prelievo* nel modo seguente: si consiglia di aggiungere l'attributo coordinate all'entità *sito*, associandolo a una coppia di coordinate, ad esempio riferite al centro geografico o all'ingresso principale dell'edificio, che costituirebbero una chiave per l'entità. Inoltre, l'entità *punto di prelievo* potrebbe essere trasformata in un'entità debole rispetto al *sito*, implicitando il vincolo imposto dall'associazione di un punto di prelievo a un sito, secondo il quale un punto di prelievo deve essere situato all'interno del perimetro del sito di cui fa parte. Al *punto di prelievo* potrebbero essere attribuite proprietà che ne descrivano la posizione all'interno del sito, come il piano, la stanza o il tipo di componente idraulico, da cui è stato prelevato il campione.

Complessivamente, gli adeguamenti proposti esercitano un impatto positivo sulla gestione dei vincoli di integrità del database, poiché risultano logicamente più immediati e più facili da implementare rispetto alle soluzioni precedenti, e contribuiscono a fornire una visione ordinata e completa dei dati relativi alla diffusione della Legionella.

2.c.I. Diagramma E-R che raccoglie le modifiche proposte

A seguito di queste considerazioni, si propone una revisione dello schema. La nuova versione è modellata secondo la notazione classica E-R[14] che consente di rappresentare in modo chiaro e conciso le entità, le relazioni e gli attributi del database.

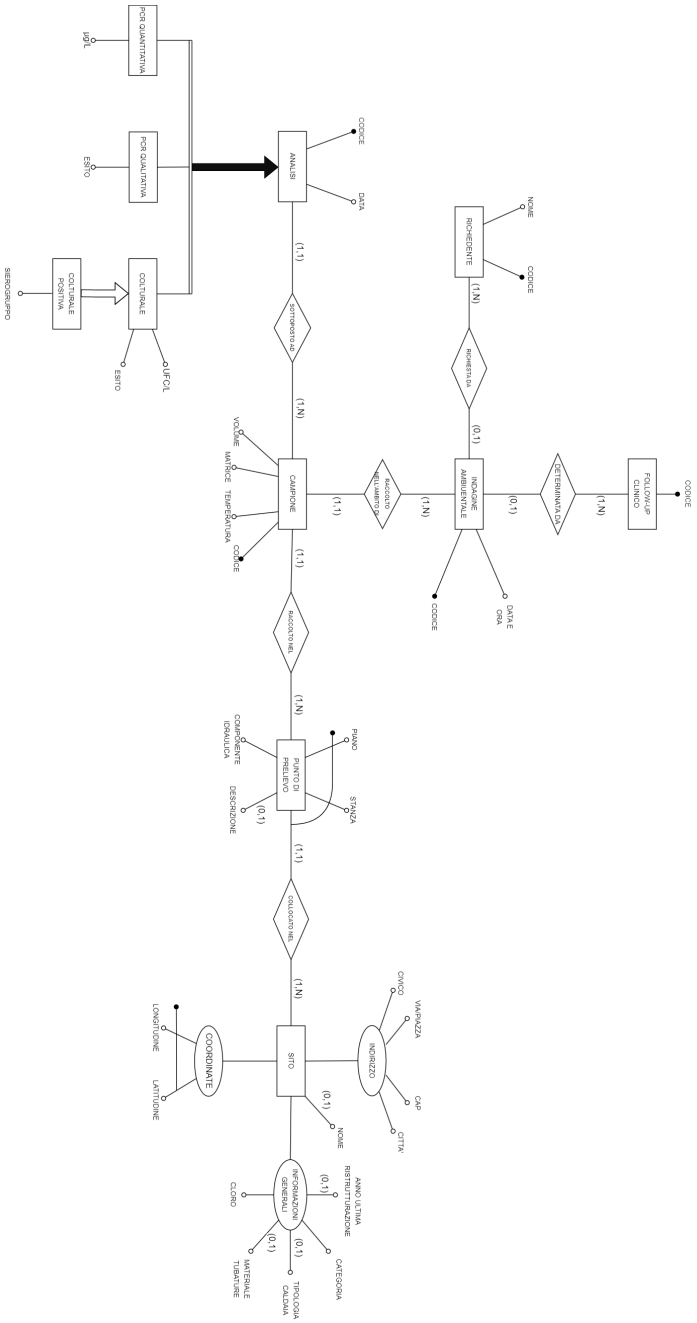

Figura 2: Diagramma ER

[14][4] P. Atzeni, S. Ceri, S. Paraboschi, e R. Torlone, *Database Systems: Concepts, Languages & Architectures*. 1999.

3. Integrazione dei nuovi requisiti nella base di dati

Come accennato in precedenza, la progettazione concettuale della base di dati deve essere adeguata alle nuove esigenze emerse a seguito dei colloqui con i ricercatori di ARPA FVG. In questa sezione si procede con l'integrazione dei nuovi requisiti nella base di dati, partendo dallo schema concettuale proposto in conclusione del capitolo precedente.

3.a. Requisiti e proposte di modfifica dello schema

Le nuove informazioni sono finalizzate a rendere la base di dati più completa e funzionale. In particolare, è stata considerata l'opportunità di introdurre ulteriori entità e attributi, allo scopo di memorizzare dati aggiuntivi relativi ai campioni raccolti nel corso delle indagini ambientali e ai siti coinvolti. Di seguito sono elencati i requisiti, non strutturati, che hanno guidato l'integrazione dei nuovi elementi e le corrispondenti proposte di modifica dello schema.

Dati meteorologici

Si ritiene opportuno mantenere le informazioni relative agli aspetti meteorologici e climatici dei siti in cui vengono condotte le indagini ambientali, poiché tali dati possono essere utili per valutare l'impatto delle condizioni ambientali sulla diffusione del batterio e per individuare eventuali correlazioni tra la presenza di Legionella e particolari fattori climatici. Tali informazioni sono raccolte presso le stazioni meteorologiche presenti sul territorio e comprendono dati relativi a temperatura, umidità e pressione atmosferica. Nella base di dati si propone di introdurre un'entità denominata *stazione meteorologica*, identificata dalla posizione geografica, che può essere rappresentata attraverso l'indirizzo oppure le coordinate, e che conserva i dati meteorologici raccolti. Questa entità è associata alla tabella *sito* nel seguente modo: ogni sito è in relazione con la stazione meteorologica più vicina, la quale forsnisce i dati relativi alle condizioni climatiche del luogo.

Analisi del pH

Una seconda considerazione riguarda l'opportunità di ampliare il campo di azione delle analisi condotte sui campioni prelevati durante le indagini ambientali. In particolare, si suggerisce di introdurre un nuovo tipo di analisi, denominata *analisi del pH*, volta a misurare il livello di acidità o alcalinità dell'acqua campionata. Questo parametro è di fondamentale importanza per valutare la qualità dell'acqua e la presenza di Legionella, poiché il batterio prospera in acque con pH neutro o leggermente alcalino.

Informazioni genomiche

Sempre in relazione alle analisi condotte sui campioni, durante i colloqui è emerso il proposito di memorizzare le informazioni genomiche relative al batterio. In particolare, si intende raccogliere dati sulla presenza, o assenza, di specifici geni e individuare i fattori genetici che influenzano la diffusione del batterio. A tale scopo, è necessario eseguire un'analisi genomica sui campioni prelevati per identificare la sequenza del DNA di Legionella. Questa informazione è memorizzata in un'entità *analisi genomica*, che rappresenta una specializzazione dell'entità *analisi*, e contiene l'intera sequenza del DNA di Legionella, espressa mediante le quattro lettere che indicano le basi azotate (A, T, C, G).

A ciascun genoma sequenziato si intende associare i geni noti di Legionella, presenti nei database di riferimento di BLAST[15] corrispondenti. Tali geni sono memorizzati in un'entità *gene*, identificata univocamente mediante una chiave corrispondente al relativo protein ID[16] e caratterizzata dal nome del gene, se presente nel database utilizzato per l'analisi. A questa entità, che ha lo scopo di conservare informazini stabili e ben definite sui geni noti di Legionella, si propone di associare un'entità *gene del genoma*, che rappresenta i geni individuati per ogni genoma sequenziato. Si tiene traccia, tramite i principali parametri restituiti dalle query BLAST, del fattore di similarità tra i geni noti e quelli individuati tramite l'analisi. Questo approccio ha lo scopo di consentire, in futuro, a seguito del progresso delle tecniche di riconoscimento genetico e dell'aumento dei dati disponibili, una rivalutazione dei geni identificati, al fine di determinare se emergono geni con maggiore somiglianza rispetto a quelli attualmente presenti nel genoma analizzato. Ogni entry della tabella *gene di un genoma* è distinta dall'insieme formato dal protein-ID, dal genoma di cui è parte e dalla posizione assoluta all'interno del genoma.

Infine, si intende registrare per ciascun *gene del genoma* la sua posizione relativa rispetto agli altri geni all'interno del profilo genetico sequenziato. Questa

[15]Basic Local Alignment Search Tool. https://blast.ncbi.nlm.nih.gov/Blast.cgi

[16]Identificativo univoco associato a ciascuna proteina mappata nei database di riferimento di BLAST.

informazione è essenziale per valutare la prossimità tra i geni e per identificare eventuali pattern di distribuzione specifici all'interno del genoma di Legionella. In termini pratici, si suggerisce l'introduzione di una relazione auto-referenziale che coinvolga l'entità *gene del genoma*, al fine di stabilire un legame tra i geni identificati e la loro posizione, relativa[17] all'interno del genoma sequenziato. La cardinalità di tale relazione sarà definita come (0,1) a (0,1), indicando che ogni gene può essere associato a zero o un gene rispetto al quale è sequenziale nel profilo genetico. Questa configurazione tiene conto della limitata conoscenza attuale sui geni di Legionella, che potrebbe comportare l'assenza di associazioni per alcune aree del genoma. Si noti che la relazione di sequenzialità tra i geni è monodirezionale, ovvero è conservata, per ogni gene, solamente l'informazione relativa al predecessore nel profilo genetico. Tale scelta è dettata dalla proposizione di mantenere basso il costo computazionale per la gestione delle informazioni, evitando così il rischio di inconsistenza dovuto alla duplicazione dei dati. In questo modo si elude l'introduzione di vincoli di integrità aggiuntivi, preferendovi piuttosto l'aumento della complessità di un' eventuale query finalizzate a ottenere l'informazione nel senso opposto a quello definito dalla relazione. Si osserva che gli unici vincoli di integrità che si rendono necessari sono i seguenti: un gene del genoma non può essere associato a se stesso, né può essere associato a un altro gene se esistono geni noti che hanno posizione assoluta maggiore rispetto al gene con il quale si vuole stabilire la relazione di sequenzialità, ma minore rispetto al gene inserito.

[17]Definita in relazione alla prossimità ad altri geni all'interno del genoma sequenziato.

3.b. Diagramma E-R integrante le nuove esigenze

A seguito delle modifiche proposte, è realizzato il seguente diagramma E-R.

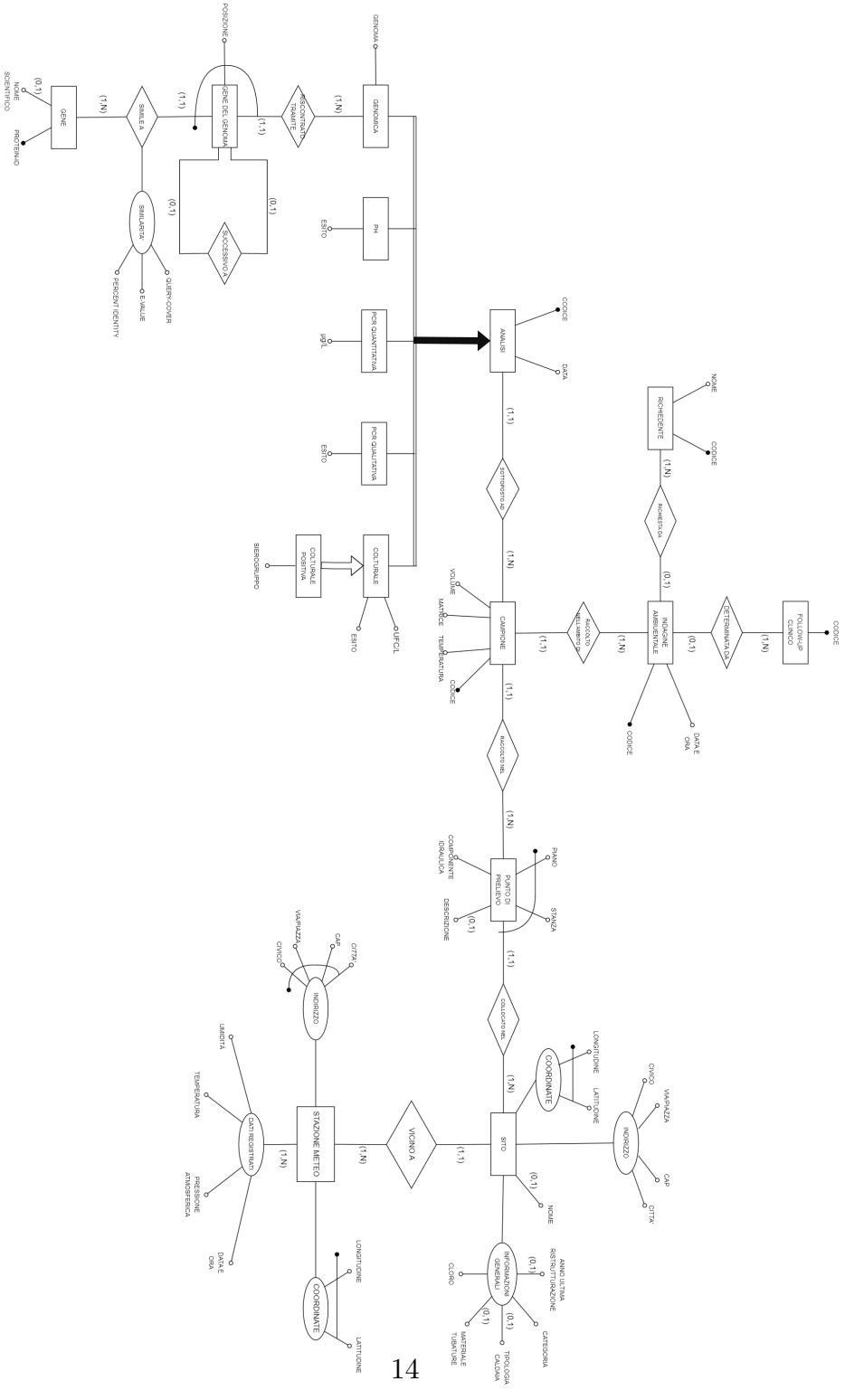

Figura 3: Diagramma ER

3.B.I. Note

Si noti come che l'introduzione di nuove entità e relazioni, sebbene arricchisca il quadro di informazioni memorizzate nel database, comporta un forte aumento della complessità del sistema. Più specificamente, l'aggiunta delle entità coinvolte nella memorizzazione delle informazioni genomiche richiede maggiore attenzione, in quanto, per garantire la coerenza dei dati con le informazioni disponibili nei database di BLAST o degli altri strumenti che possono essere utilizzati per l'individuazione e la classificazione dei geni, è necessario aggiornare costantemente le istanze dell'entità *gene* con i dati più recenti.

In ultimo si segnala che le principali operazioni eseguite sulla base di dati riguardano l'inserimento, la modifica e la cancellazione dei dati. Al contrario, le operazioni di interrogazione sono limitate a un numero ristretto di query, finalizzate a ottenere informazioni di tipo spaziale sui campioni, sulle analisi effettuate e sui risultati ottenuti oppure informazioni genetiche. Pertanto, si preferisce adottare una struttura facilmente manutenibile e ottimizzata per le operazioni fondamentali, che risulta già adeguata per l'esecuzione delle operazioni sopra menzionate, piuttosto che una struttura più complessa, progettata per ottimizzare le interrogazioni, ma che comporterebbe un costo maggiore per la gestione dei dati.

Le considerazioni relative ai vincoli di integrità sono posticipate al capitolo successivo, nel quale, terminata la fase di progettazione, sarà possibile ottenere una visione del tutto trasparente e definitiva delle entità coinvolte nel sistema dei relativi attributi e delle relazioni tra di esse.

4. Progettazione logica della base di dati

4.a. Ristrutturazione del modello concettuale: Semplificazione delle generalizzazioni e degli attributi composti

Ultimata la fase di progettazione concettuale della base di dati, è opportuno effettuare un'ultima revisione del modello al fine di elaborarne la struttura finale, priva di elementi discrezionali. In questa unità sono riportate le modifiche, congiuntamente alle motivazioni che le guidano, apportate allo schema E-R proposto al paragrafo 3.b, con l'obbiettivo di risolvere generalizzazioni, attributi composti e attributi multivalori presenti in figura.

In prima istanza sono trattati gli aspetti riguardanti le entità coinvolte nelle relazioni di generalizzazione.

Per quanto concerne l'entità *analisi*, si propone di rimuovere, in tutto il suo complesso, la relazione di generalizzazione associando piuttosto le differenti tipologie di analisi ai campioni su cui sono eseguite. Ognuna delle analisi è identificata in modo esclusivo da un codice e caratterizzata sia della data di esecuzione, proprietà ereditata dall'entità soppressa *analisi*, sia dagli attributi caratterizzanti di ciascuna specializzazione. Inoltre, pur non essendo del tutto rigoroso dal punto di vista scientifico, si propone di riassumere le tabelle rappresentative delle analisi *PCR qualitativa* e *PCR quantitativa* in un'unica soluzione denominata *analisi PCR*, che conserva informazioni di entrambe le nature sui campioni analizzati. Questa semplificazione è ritenuta lecita in quanto i risultati prodotti dalle due analisi sono intrinsecamente correlati e possono essere memorizzati in modo più efficiente all'interno di un'unica entità. Infatti la PCR qualitativa, che rileva la presenza del DNA di Legionella, è, nell'ipotesi in cui restituisce un risultato positivo, seguita dala PCR quantitativa, la quale misura la concentrazione del batterio nel campione. La soluzione proposta consente di alleggerire la struttura del database e di semplificare le operazioni di inserimento e consultazione dei dati, senza introdurre perdite di informazioni

né comportare un'aumento della complessità del sistema inteso come l'introduzione di vincoli di integrità. Si noti che non tutte le analisi sono eseguite su tutti i campioni, ma, talvolta, solo su una parte di essi. Ad esempio, come suggerito dalle linee guida per la prevenzione ed il controllo della legionellosi, «poichè la q-PCR è effettivamente vantaggiosa per molteplici aspetti ma non ancora validata a livello internazionale, essa può, ad oggi, essere solo consigliata per una rapida analisi di numerosi campioni prelevati da siti probabilmente associati ad un caso o ancor più a un cluster di legionellosi, potendo in tempi brevi escludere i siti negativi ed identificare quelli positivi»[18]. In altre parole, le stesse linee guida suggeriscono di eseguire l'analisi colturale solo in caso di risultato positivo alla q-PCR, senza tuttavia stabilire una convenzione. Tale mancanza consente diverse interpretazioni e, pertanto, si è deciso di definire la cardinalità della relazione tra *campione* e le diverse *analisi* come "(0,1) a (1,1)", stabilendo che un campione può essere associato a zero o una sola analisi specifica, mentre ogni analisi è sempre associata a un campione. Questa scelta garantisce anche la retrocompatibilità del sistema. Infatti non essendo possibile associare a campioni già esistenti le analisi aggiunte successivamente, ovvero *analisi del pH* e *analisi genomica*, è opportuno che i campioni non siano obbligatoriamente associati a tutte le tipologie di analisi.

Per quanto concerne la specializzazione relativa all'*analisi colturale*, ovvero l'*analisi colturale positiva*, se ne suggerisce la sostituzione con un attributo denominato sierogruppo, che è proprio dell'entità *analisi colturale*. Tale modifica permette di conservare le informazioni relative al sierogruppo di Legionella identificato nel campione, senza introdurre una nuova entità e risparmiando dunque spazio. Si noti che questa modifica implica l'introduzione di un vincolo di integrità che assicuri che solamente le analisi colturali positive siano associate a un sierogruppo. Maggiori dettagli sui vincoli di integrità saranno forniti nel capitolo successivo.

Un'ulteriore elemento di riflessione riguarda la risoluzione degli attributi composti.

In riferimento alla relazione *simile a* che coinvolge le entità *gene* e *gene del genoma*, si propone di scomporre l'attributo similarità nelle tre componenti che lo costituiscono: percent identity, query-cover e e-value. Inoltre, la cardinalità della relazione suggerisce di ricollocare questi attributi in modo tale che siano riferiti all'entità *gene del genoma*.

In merito all'attributo composto e multivalotre *dati registrati*, afferente all'entità *stazione meteorologica*, a soluzione più adeguata consiste nella sua sostituzione con una nuova tabella denominata *dati meterelogici* associata tramite una relazione del tipo 1 a N all'entità *stazione meteorologica*. Questa nuova

[18][1], *Linee guida per la prevenzione ed il controllo della legionellosi*, p. 21

entità è debole rispetto alla *stazione meteorologica* e conserva le seguenti informazioni: la data di acquisizione di ogni set di dati, che costituisce parte della chiave primaria, la temperatura, l'umidità e la pressione atmosferica.

In conclusione, relativamente agli attributi indirizzo e coordinate, si ritiene opportuno adottare la medesima soluzione per tutte le apparizioni nello schema: l'attributo coordinata è sostituito dalla coppia di latitudine e longitudine, che costituiscono la chiave primaria delle entità cui sono riferite, ovvero *sito* e *stazione meteorologica*;l'attributo indirizzo è risolto in modo analogo, sostituendolo con i campi via, numero civico, CAP e città. Questa soluzione è dettata dall'intenzione di sfruttare le informazioni geografiche per condurre analisi mirate sulla diffusione della Legionella e per identificare eventuali cluster. In particolare, la soluzione proposta consente di semplificare notevolmente la struttura delle interrogazioni necessarie per ottenere tali informazioni.

Da sistemare: Ad esempio, l'implementazione della query Sezione 5.c.II.I di ricerca di tutti i campioni positivi prelevati in una certa data in un certa via di una certa città, potrebbe restituire risultati inesatti qualora l'indirizzo fosse un'unica stringa di testo, perchè non sarebbe possibile distinguere gli elementi via e città. La soluzione proposta, invece, permette di ottenere risultati corretti e coerenti con le aspettative.

4.b. Diagramma E-R finale

Il diagramma E-R finale, risultante dalla rielaborazione dello schema proposto al paragrafo 3.b, è presentato di seguito.

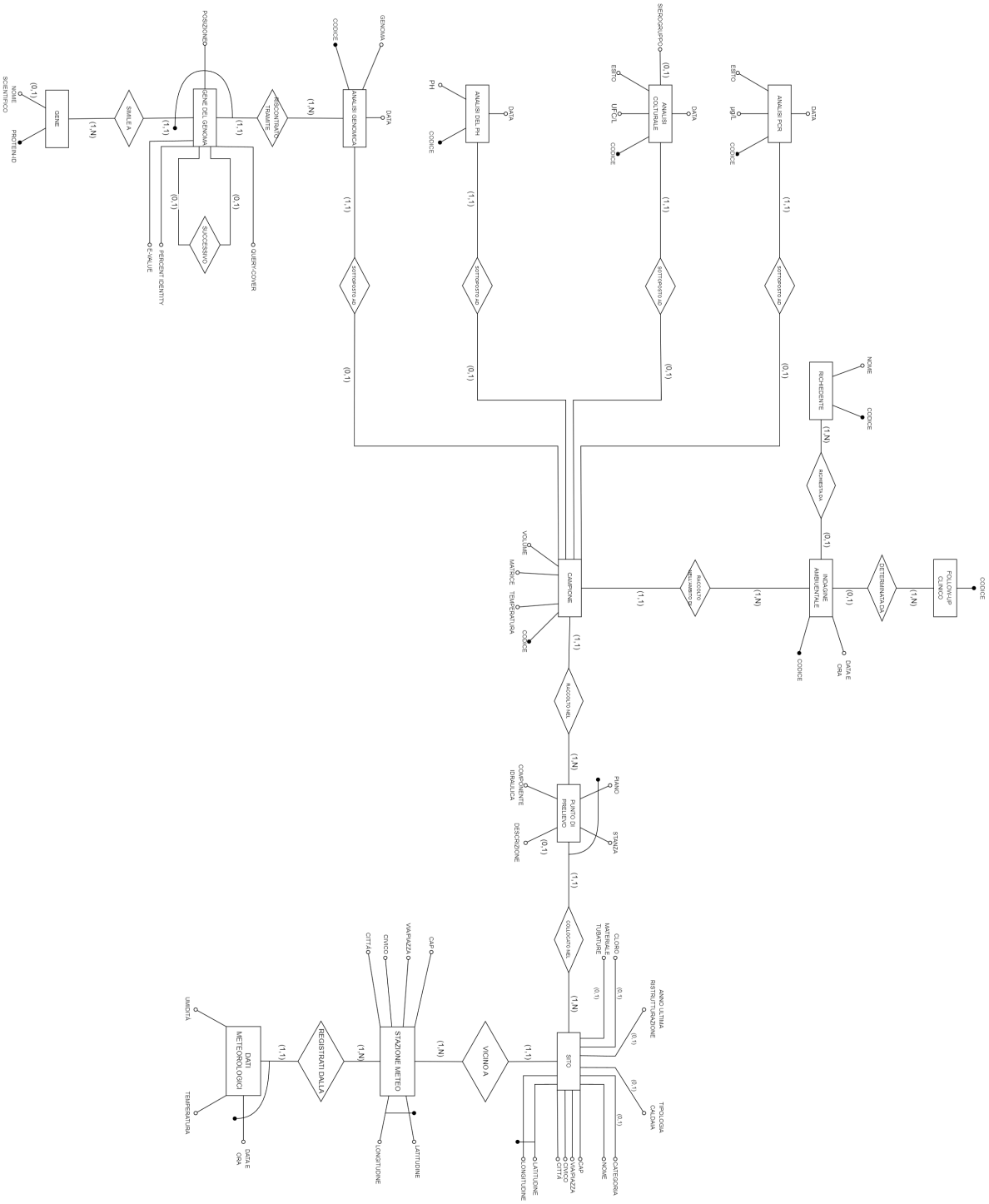

Figura 4: Diagramma ER

4.c. Illustrazione delle decisioni di progettazione nella trasformazione dal modello concettuale a quello logico

La trasposizione del modello concettuale in quello logico comporta la definizione delle tabelle, dei rispettivi campi, e quindi dei domini, delle chiavi primarie e delle modalità di associazione tra le tabelle. In questa sezione vengono presentate le scelte progettuali adottate.

La traduzione delle entità in tabelle è diretta e non comporta particolari difficoltà. Ogni entità, infatti, è rappresentata mediante una matrice in cui ogni attributo corrisponde a una colonna.

Per quanto riguarda le relazioni è essenziale considerare tre tipologie fondamentali: le relazioni uno a uno, le relazioni uno a molti e le relazioni molti a molti. Le relazioni uno a uno sono le più complesse in quanto non è immediatamente chiaro quale entità debba essere scelta per mappare la relazione, ovvero in quale entità inserire la chiave esterna. Nel nostro contesto si presentano due situazioni principali: la relazione autoreferenziale tra i geni del genoma e la relazione tra i campioni e le analisi. Nel primo caso, la relazione è mappata sull'entità *gene del genoma* poichè, sebbene gli strumenti di analisi presentino alcune limitazioni, nella maggioranza dei genomi analizzati esiste un'effettiva sequenzialità tra i geni. Pertanto lo spazio sprecato a causa della mancanza di informazioni è limitato e dunque non giustifica l'introduzione di una nuova tabella che, pur limitando lo spazio utilizzato, porterebbe problemi di integrità e complessità.

Per quanto riguarda le relazioni tra campioni e analisi, si è deciso di mappare la relazione sull'entità *analisi*. Si osserva che la soluzione alternativa, ovvero quella di mappare la relazione sull'entità *campione*, potrebbe comportare perdite di spazio a causa della presenza di campioni non analizzati rispetto ad un esame specifico.

Le relazioni uno a molti, invece, sono più semplici da gestire, in quanto la chiave esterna è necessariamente inserita nell'entità che rappresenta il lato "uno" della relazione. Infine, le relazioni molti a molti sono gestite mediante l'introduzione di una tabella di associazione, che contiene le chiavi esterne delle due entità coinvolte.

Per concludere, le entità deboli, come *dati meteorologici*, *punto di prelievo* e *gene del genoma* la chiave primaria è composta dalla chiave primaria dell'entità forte, o delle entità forti, a cui sono associate e un attributo che ne identifica univocamente l'istanza all'interno dell'entità forte.

4.d. Schema relazionale

Sulla base delle considerazioni precedenti, si procede con la definizione dello schema relazionale, che rappresenta la struttura logica del database.

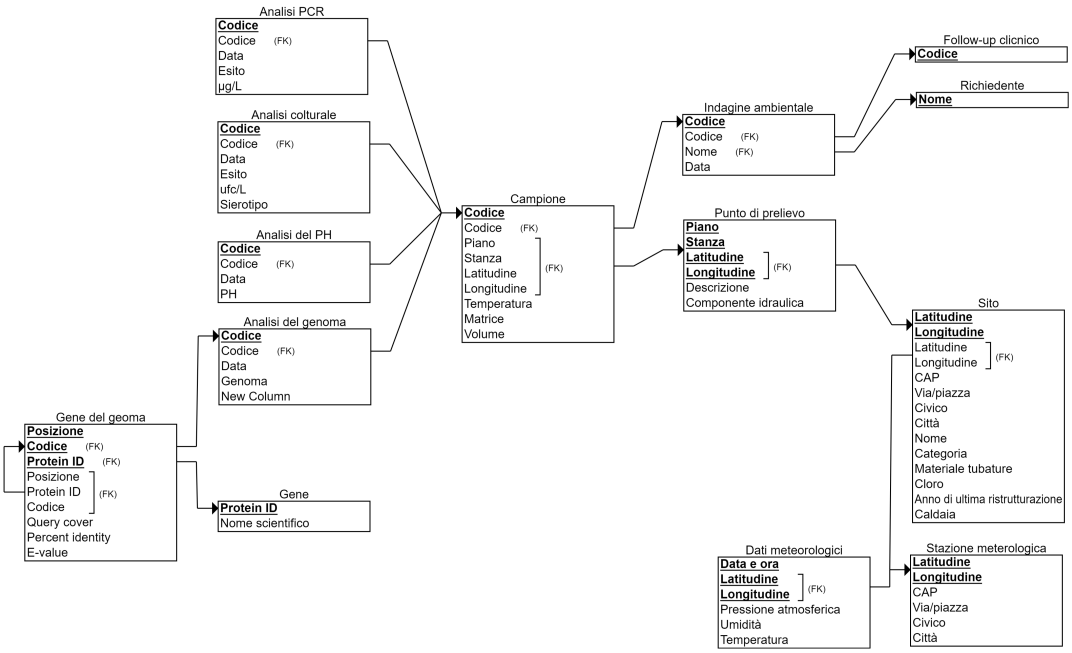

Figura 5: Schema relazionale

5. Progettazione fisica della base di dati: definizione dei domini, dei vincoli di integrità e impplementazione del codice SQL

In questa sezione viene eseguita l'implementazione della base di dati, iniziando dalla definizione dei domini, ovvero l'insieme dei valori ammissibili per ciascuna colonna, e proseguendo con la creazione delle tabelle e delle funzioni che implementano i vincoli che assicurano l'integrità dei dati.

Il DBMS scelto per la realizzazione del database è PostgreSQL, un sistema di gestione di basi di dati relazionali rilasciato per la prima volta nel 1989. La scelta è motivata dalla flessibilità ed estendibilità del sistema, che consente di implementare vincoli di integrità complessi e di gestire grandi quantità di dati in modo efficiente.

5.a. Definizione dei domini

Per prima cosa è necessario definire i vincoliLa maggior parte dei domini relativi alle colonne di ciascuna tabella è facilmente determinabile. Tuttavia, alcuni domini richiedono una definizione più dettagliata per garantire una corretta rappresentazione dei dati e facilitare l'esecuzione delle operazioni di interrogazione.

Primariamente si considerino i domini relativi alla quantificazione della Legionella nei campioni, espressi in ufc/l e µg/l. Per entrambi è opportuno ridurre il dominio ai valori interi positivi, poichè non ha senso esprimere la presenza di Legionella con valori negativi. Un caso analogo si presenta per i valori che misurano il volume di un campione, l'umidità e la pressione atmosferica, per i quali si propone di definire il dominio come un numero decimale positivo, in quanto non ha senso avere valori negativi per queste grandezze fisiche.

```sql
-- Dominio per il valore intero non negativo
CREATE DOMAIN INT_POS AS INTEGER
    CONSTRAINT valore_non_negativo CHECK (VALUE >= 0);

-- Dominio per il valore reale non negativo
CREATE DOMAIN FLOAT_POS AS FLOAT
    CONSTRAINT valore_non_negativo CHECK (VALUE >= 0);
```

In secondo luogo si analizzi il dominio relativo al parametro di misurazione del pH. Per il fatto che il range di valori ammissibili per il pH è compreso tra 0 e 14, si propone di definire il dominio del pH come un numero decimale rientrante in questo intervallo.

```
CREATE DOMAIN PH AS FLOAT
    CONSTRAINT ph_range CHECK (VALUE >= 0 AND VALUE <= 14);
```

Un ulteriore aspetto da considerare riguarda le colonne categoria e matrice relative rispettivamente alle tabelle *sito* e *campione*. Per quanto riguarda la colonna categoria, si propone di limitare il dominio a pochi vocaboli appartenenti ad un ristretto insieme semantico, come ad esempio "ospedaliero", "termale", "alberghiero", "pubblico" e "privato". Si precisa che il valore "pubblico" include tutti quegli edifici, non afferenti alle categorie specificate, destinati alla fruizione da parte di un'ampia e variegata utenza. Analogamente, per la colonna matrice, si propone di fissare un dominio che comprenda solo valori appartenenti a un insieme finito di matrici, come ad esempio "acqua", "aria" e "biofilm" e "sedimento".

```
-- Tipo enum per la categoria di un sito
CREATE TYPE CATEGORIA AS
    ENUM ('ospedaliero', 'termale', 'alberghiero', 'pubblico', 'privato');

-- Tipo enum per la matrice di un campione
CREATE TYPE MATRICE AS
    ENUM ('acqua',  'aria', 'biofilm', 'sedimento');
```

La colonna CAP delle tabelle *sito* e *stazione meteorologica* rappresenta un aggiuntivo aspetto notevole. Poichè il CAP è un codice numerico formato da cinque cifre, si suggerisce di definire un dominio di tipo integer che accetti esclusivamente valori numerici di tale lunghezza.

```
CREATE DOMAIN CAP AS INTEGER
    CONSTRAINT cap_range CHECK (VALUE >= 10000 AND VALUE <= 99999);
```

Inoltre, si propone di restringere l'intervallo dei valori ammessi per le colonne *percent-identity*, *query-cover* e *e-value* dell'entità *gene del genoma*. In termini pratici, si suggerisce di definire un dominio di tipo float compresi tra 0 e 100 per le colonne *percent-identity* e *query-cover*, in quanto rappresentano percentuali di similarità tra i geni noti e quelli individuati tramite l'analisi. Per quanto concerne l'*e-value*, invece, si propone di utilizzare il dominio FLOAT_POS definito in precedenza, in quanto si vuole rappresentare un valore numerico positivo.

```
CREATE DOMAIN PERCENT AS FLOAT
    CONSTRAINT percent_range CHECK (VALUE >= 0 AND VALUE <= 100);
```

Infine, per quanto riguarda gli attributi relativi alle coordinate geografiche, ovvero latitudine e longitudine, il dominio deve essere limitato a valori compresi tra -90 e 90 per la latitudine e tra -180 e 180 per la longitudine.

```
-- Dominio per latitudine
CREATE DOMAIN LATITUDINE AS REAL
    CONSTRAINT latitudine_range CHECK (VALUE >= -90 AND VALUE <= 90);

-- Dominio per longitudine
CREATE DOMAIN LONGITUDINE AS REAL
    CONSTRAINT longitudine_range CHECK (VALUE >= -180 AND VALUE <= 180);
```

5.A.I. Note

Sulla base delle osservazioni riportate in diversi articoli scientifici riguardanti lo studio degli aspetti genetici del batterio, come ad esempio *Draft genome sequences from 127 Legionella spp. strains isolated in water systems linked to legionellosis outbreaks*[19], è emerso che la lunghezza media del genoma di Legionella pneumophila è di circa 3.500.000 coppie di basi, con una significativa variabilità tra i genomi sequenziati.

In considerazione alle dimensioni dell'oggetto, si propone di assegnare un dominio di tipo text[20]. Questo tipo di dato consente la memorizzazione di stringhe di lunghezza arbitraria, risultando particolarmente adatto per la conservazione di sequenze di DNA.

Si potrebbe anche impostare un limite superiore e usare varchar, ma sarebbe meglio aggiungere uno step di conferma per valori superiori a 7.000.000 bp (il doppio rispetto alla media, ovvero un possibile errore di duplicazione dei dati) in un'eventuale interfaccia per interagire con il database per lasciare flessibilità a questo livello

Si evidenzia inoltre che, in termini di occupazione della memoria, la politica TOAST[21] tipica di PostgreSQL consente una gestione efficiente anche per attributi di grandi dimensioni, allocando i dati in pagine separate e comprimendoli per ridurre lo spazio complessivo occupato.

[19] [5] A. Colautti *et al.*, «Draft genome sequences from 127 Legionella spp. strains isolated in water systems linked to legionellosis outbreaks», *Microbiol Resour Announc*, vol. 13, fasc. 6, 2024, doi: 10.1128/mra.01154-23.

[20] https://www.postgresql.org/docs/current/datatype-character.html

[21] https://www.postgresql.org/docs/current/storage-toast.html

5.b. Creazione delle tabelle

Il codice per la creazione delle tabelle è banalmente ottenuto dal modello relazionale. Tuttavia, merita particolare attenzione la gestione dei vincoli di chiave esterna. In particolare, è necessario considerare il comportamento delle chiavi esterne nei casi di eliminazione o aggiornamento di una riga a cui queste fanno riferimento. In questo ambito vi sono tre principali opzioni, ovvero l'impedimento dell'operazione (RESTRICT), che comporta il rifiuto dell'operazione stessa, l'azione di cascata (CASCADE), che comporta l'aggiornamento o la cancellazione delle righe collegate alla riga interessata, e l'assegnazione di un valore nullo (SET NULL), che imposta il valore nullo nelle righe che fanno riferimento alla riga eliminata o modificata.

Per ragioni di spazio vengono forniti alcuni esempi di creazione delle tabelle, mentre il codice completo è riportato in appendice.

Un aspetto rilevante riguarda la cancellazione di un campione. In generale si ritiene opportuno di eliminare i dati associate al campione, poiché perderebbero di significato in sua assenza. Tuttavia, si propone di impedire l'operazione di cancellazione qualora il campione sia associato ad un'analisi del genoma. Tale decisione è finalizzata a interrompere la catena di eliminazione, che coinvolgerebbe tutte le informazioni relative ai dati genomici osservati, al fine di evitare l'eliminazione accidentale di una grande quantità di dati. Si osserva che, per eliminare un campione, sarà sufficiente rimuovere preventivamente l'eventuale analisi genomica associata, dopodiché sarà possibile procedere con la cancellazione del campione stesso. A titolo di esempio si riportano le tabelle *analisi PCR* e *analisi del genoma*

```sql
-- Analisi PCR
CREATE TABLE Analisi_PCR (
    codice CHAR(6) NOT NULL,
    codice_campione CHAR(6) NOT NULL,
    data_ora DATE NOT NULL,
    esito BOOLEAN NOT NULL,
    µg_l INT_POS NOT NULL,

    PRIMARY KEY (codice),

    FOREIGN KEY (codice_campione) REFERENCES Campione(codice)
        ON DELETE CASCADE
        ON UPDATE CASCADE
);

-- Analisi genomica
CREATE TABLE Analisi_genomica (
    codice CHAR(6) NOT NULL,
```

```
    codice_campione CHAR(6) NOT NULL,
    data_ora DATE NOT NULL,
    genoma TEXT NOT NULL,

    PRIMARY KEY (codice),

    FOREIGN KEY (codice_campione) REFERENCES Campione(codice)
        ON DELETE RESTRICT
        ON UPDATE CASCADE
);
```

Un secondo caso di particolare rilevanza riguarda la tabella *gene del genoma*. Per quanto concerne la relazione con i geni noti di Legionella, si propone di propagare l'aggiornamento della chiave protein-ID a cascata e di impedire la cancellazione dei geni noti per i quali esistano geni del genoma associati. Relativamente alla relazione con i genomi di Legionella sequenziati, invece, è opportuno eseguire sia le operazioni di aggiornamento che di cancellazione a cascata. Tale comportamento è motivato dal fatto che, in caso di modifica di un genoma, è necessario aggiornare i geni del genoma associati; mentre, in caso di cancellazione del genoma, risulta opportuno eliminare anche i geni del genoma collegati, per evitare inconsistenze dovute alla presenza di dati orfani. Infine, riguardo alla relazione di sequenzialità tra i geni del genoma, la soluzione è immediatamente determinata dalla cardinalità della relazione: sostituire i dati registrati con il valore NULL.

```
CREATE TABLE Gene_del_genoma (
    posizione INTEGER NOT NULL,
    codice_genoma CHAR(6) NOT NULL,
    protein_ID CHAR(6) NOT NULL,
    posizione_predecessore INTEGER,
    codice_genoma_predecessore CHAR(6),
    protein_ID_predecessore CHAR(6),
    query_cover PERCENT NOT NULL,
    percent_identity PERCENT NOT NULL,
    e_value FLOAT_POS NOT NULL,

    PRIMARY KEY (posizione, codice_genoma, protein_ID),

    FOREIGN KEY (codice_genoma) REFERENCES Analisi_genomica(codice)
        ON DELETE CASCADE
        ON UPDATE CASCADE,
    FOREIGN KEY (protein_ID) REFERENCES Gene(protein_ID)
        ON DELETE RESTRICT
        ON UPDATE CASCADE,
    FOREIGN KEY (posizione_predecessore, codice_genoma_predecessore,
protein_ID_predecessore) REFERENCES Gene_genoma(posizione, codice_genoma,
protein_ID)
        ON DELETE SET NULL
        ON UPDATE CASCADE
);
```

In ultimo, si riporta una nota riguardante la tabella sito. Poiché si desidera collegare ciascun sito alla stazione meteorologica più vicina, le operazioni di aggiornamento e cancellazione delle stazioni sono risolte mediante l'applicazione di un trigger, che associa automaticamente tutti i siti legati alla stazione interessata alla stazione meteorologica più vicina. Di conseguenza, non è necessario definire un vincolo di chiave esterna per tale tabella. Una trattazione più approfondita sarà fornita nei capitoli successuvi.

```
CREATE TABLE Sito (
    latitudine LATITUDINE NOT NULL,
    longitudine LONGITUDINE NOT NULL,
    latitudine_stazione LATITUDINE NOT NULL,
    longitudine_stazione LONGITUDINE NOT NULL,
    CAP CAP NOT NULL,
    via_piazza VARCHAR(25) NOT NULL,
    civico INTEGER NOT NULL,
    città VARCHAR(25) NOT NULL,
    nome VARCHAR(25),
    categoria CATEGORIA NOT NULL,
    materiale_tubature VARCHAR(25),
    cloro BOOLEAN NOT NULL,
    anno_ultima_ristrutturazione DATE,
    caldaia VARCHAR(25),

    PRIMARY KEY (latitudine, longitudine),

    FOREIGN KEY (latitudine_stazione, longitudine_stazione) REFERENCES
Stazione_meterologica(latitudine, longitudine)
);
```

5.c. Definizione dei vincoli

A questo punto si dispone di una visione completa e definitiva della struttura del database, che rende possibile analizzare le criticità non risolte dallo schema attuale. In questa sezione sono presentati i vincoli di integrità necessari per garantire la consistenza dei dati all'interno del database, insieme alle motivazioni che ne determinano l'introduzione.

5.c.I. Vincoli di chiave esterna della tabella *Sito*

In questo paragrafo si affrontano le problematiche relative alla cancellazione e all'aggiornamento di una stazione meteorologica, come accennato nel capitolo precedente. Poiché si desidera associare ciascun sito alla stazione meteorologica più vicina, si propone di implementare un trigger che, in caso di cancellazione o aggiornamento di una stazione meteorologica, assegni automaticamente a tutti i siti precedentemente collegati alla stazione interessata la stazione meteorologica più vicina. Questa soluzione evita l'utilizzo di un vincolo RESTRICT, il

quale renderebbe più complessa la gestione delle operazioni di cancellazione e aggiornamento.

In dettaglio, il trigger di aggiornamento si occupa di aggiornare la stazione meteorologica associata a ciascun sito, sostituendola con quella più vicina in seguito alla modifica delle coordinate, o all'inserimento, di una stazione meteorologica. Il trigger di cancellazione, invece, impedisce l'eliminazione totale delle stazioni meteorologiche, assicurando che almeno una stazione meteorologica sia associata a ciascun sito. Se la cancellazione è possibile, il trigger aggiorna le coordinate riferite alle stazioni meteorologiche dei siti coinvolti con quelle degli osservatori più vicini a ciascuno. In entrambi i casi, la distanza tra le i centri meteorologici e i siti viene calcolata utilizzando la formula di Haversine, che consente di determinare la distanza tra due punti sulla superficie di una sfera, come la Terra, conoscendone le coordinate geografiche.

```
-- Funnzione di Haversine
CREATE OR REPLACE FUNCTION distance(lat1 LATITUDINE, lon1 LONGITUDINE, lat2
LATITUDINE, lon2 LONGITUDINE)
RETURNS FLOAT LANGUAGE plpgsql AS
$$
DECLARE
    R FLOAT := 6371; -- Raggio medio della Terra in km
    phi1 FLOAT := RADIANS(lat1);
    phi2 FLOAT := RADIANS(lat2);
    delta_phi FLOAT := RADIANS(lat2 - lat1);
    delta_lambda FLOAT := RADIANS(lon2 - lon1);
    a FLOAT;
    c FLOAT;
BEGIN
    a := SIN(delta_phi / 2) * SIN(delta_phi / 2) + COS(phi1) * COS(phi2) *
SIN(delta_lambda / 2) * SIN(delta_lambda / 2);
    c := 2 * ATAN2(SQRT(a), SQRT(1 - a));
    RETURN R * c;
END;
$$;

-- Trigger insert/update on Stazione_meteorologica
CREATE OR REPLACE FUNCTION update_stazione_meteorologica()
RETURNS TRIGGER LANGUAGE plpgsql AS $$
DECLARE
    lat_sito FLOAT;
    lon_sito FLOAT;
    lat_stazione FLOAT;
    lon_stazione FLOAT;
    current_distance FLOAT;
    min_distance FLOAT;
    lat_stazione_vicina FLOAT;
    lon_stazione_vicina FLOAT;
BEGIN
    -- Aggiorna i siti per riflettere la stazione meteorologica più vicina
    FOR lat_sito, lon_sito IN
        SELECT lat, lon
        FROM Sito
```

```
    LOOP
        min_distance := 'infinity'; -- Inizializzo la distanza minima a infinito
        FOR lat_stazione, lon_stazione IN
            SELECT lat, lon
            FROM Stazione_meteorologica
        LOOP
            current_distance := calculate_distance(lat_sito, lon_sito,
                                                   lat_stazione, lon_stazione);
            IF current_distance < min_distance THEN
                min_distance := current_distance;
                lat_stazione_vicina := lat_stazione;
                lon_stazione_vicina := lon_stazione;
            END IF;
        END LOOP;

        -- Aggiorna il sito con la stazione meteorologica più vicina
        UPDATE Sito
        SET latitudine_stazione_meteorologica = lat_stazione_vicina,
            longitudine_stazione_meteorologica = lon_stazione_vicina
        WHERE lat = lat_sito AND lon = lon_sito;
    END LOOP;

    RETURN NEW;
END;
$$;

CREATE TRIGGER update_stazione_meteorologica_trigger
AFTER INSERT OR UPDATE ON Stazione_meteorologica
FOR EACH ROW
EXECUTE FUNCTION update_stazione_meteorologica();

-- Trigger delete on Stazione_meteorologica
CREATE OR REPLACE FUNCTION delete_stazione_meteorologica()
RETURNS TRIGGER LANGUAGE plpgsql AS $$
DECLARE
    lat_sito FLOAT;
    lon_sito FLOAT;
    lat_stazione FLOAT;
    lon_stazione FLOAT;
    current_distance FLOAT;
    min_distance FLOAT;
    lat_stazione_vicina FLOAT;
    lon_stazione_vicina FLOAT;
BEGIN

    -- Controllo che ci siano stazioni meteorologiche rimaste
    IF (SELECT COUNT(*) FROM Stazione_meteorologica) = 1 THEN
        RAISE EXCEPTION 'Non è possibile eliminare tutte le stazioni
meteorologiche.';
        RETURN OLD;
    END IF;

    -- Aggiorna i siti per riflettere la stazione meteorologica più vicina rimasta
    FOR lat_sito, lon_sito IN
        SELECT lat, lon
        FROM Sito
        WHERE latitudine_stazione_meteorologica = OLD.lat AND
              longitudine_stazione_meteorologica = OLD.lon
```

```
    LOOP
        min_distance := 'infinity'; -- Inizializzo la distanza minima a infinito
        FOR lat_stazione, lon_stazione IN
            SELECT lat, lon
            FROM Stazione_meteorologica
        LOOP
            current_distance := calculate_distance(lat_sito, lon_sito,
                                                   lat_stazione, lon_stazione);
            IF current_distance < min_distance THEN
                min_distance := current_distance;
                lat_stazione_vicina := lat_stazione;
                lon_stazione_vicina := lon_stazione;
            END IF;
        END LOOP;

        -- Aggiorna il sito con la nuova stazione meteorologica più vicina
        UPDATE Sito
        SET latitudine_stazione_meteorologica = lat_stazione_vicina,
            longitudine_stazione_meteorologica = lon_stazione_vicina
        WHERE lat = lat_sito AND lon = lon_sito;
    END LOOP;

    RETURN OLD;
END;
$$;

CREATE TRIGGER delete_stazione_meteorologica_trigger
BEFORE DELETE ON Stazione_meteorologica
FOR EACH ROW
EXECUTE FUNCTION delete_stazione_meteorologica();
```

5.c.II. Vincoli relativi ai dati

Si consideri, in primo luogo, l'entità *analisi colturale*. La corretta formazione dei dati registrati a seguito di ciascuna analisi prevede l'introduzione dei seguenti vincoli di integrità relativi ai casi di positività del campione: ad ogni campione positivo deve essere associato un sierogruppo di Legionella, ovvero quello individuato dall'analisi, mentre ad ogni campione negativo non deve essere associato alcun sierogruppo; ad ogni campione positivo deve essere associato un valore di unità formanti colonia su litro (ufc/l) maggiore di zero, mentre ad ogni campione negativo deve essere associato il valore zero.

Per quanto riguarda l'entità *analisi PCR*, è individuato il seguente vincolo: ad ogni campione positivo deve essere associato un valore di microgrammi su litro (µg/l) maggiore di zero, mentre ad ogni campione negativo deve essere associato il valore zero.

Per l'entità *analisi del pH*, è opportuno introdurre una restrizione che garantisca che il valore del pH sia compreso tra 0 e 14, parametri che definiscono il range di valori ammissibili per il pH.

Un ulteriore accorgimento deve essere impiegato nel caso dei campioni. Infatti, come già accenntato, poichè un'indagine ambientale è una collezione di

campioni raccolti in una stessa data, in un sito specifico, è necessario garantire che tutti i campioni associati a un'indagine siano prelevati nello stesso sito.

Infine, per quanto riguarda l'entità *gene del genoma* è necessario fare alcune considerazioni sulla relazione di sequenzialità tra i geni. In particolare, si propone di introdurre diversi vincoli che assicurino che a un gene di un genoma non possa essere associato un gene di un genoma diverso né se stesso, né possa essere associato a un altro gene dello stesso genoma, qualora esistano altri geni con posizione assoluta maggiore rispetto al gene con cui si intende stabilire la relazione di sequenzialità, ma minore rispetto al gene considerato. Questo vincolo è necessario per garantire la corretta rappresentazione della sequenza genetica di Legionella e per evitare situazioni di inconsistenza.

5.c.II.I. query sui campioni positivi in una certa data, in una certa via di una certa città

6. Appendice

6.a. Codice SQL per la creazione delle tabelle

```sql
-- Stazione meteorologica
CREATE TABLE Stazione_meteorologica (
    latitudine LATITUDINE NOT NULL,
    longitudine LONGITUDINE NOT NULL,
    via VARCHAR(25) NOT NULL,
    numero_civico INTEGER NOT NULL,
    CAP CAP NOT NULL,
    città VARCHAR(25) NOT NULL,

    PRIMARY KEY (latitudine, longitudine)
);

-- Dati meteorologici
CREATE TABLE Dati_meteorologici (
    data_ora DATETIME,
    latitudine_stazione LATITUDINE NOT NULL,
    longitudine_stazione LONGITUDINE NOT NULL,
    temperatura FLOAT NOT NULL,
    umidità FLOAT_POS NOT NULL,
    pressione_atmosferica FLOAT NOT NULL,

    PRIMARY KEY (data_ora, latitudine_stazione, longitudine_stazione),

    FOREIGN KEY (latitudine_stazione, longitudine_stazione) REFERENCES
Stazione_meterologica(latitudine, longitudine)
        ON DELETE CASCADE
        ON UPDATE CASCADE
);

-- Sito
CREATE TABLE Sito (
    latitudine LATITUDINE NOT NULL,
    longitudine LONGITUDINE NOT NULL,
    latitudine_stazione LATITUDINE NOT NULL,
    longitudine_stazione LONGITUDINE NOT NULL,
    CAP CAP NOT NULL,
    via_piazza VARCHAR(25) NOT NULL,
    civico INTEGER NOT NULL,
    città VARCHAR(25) NOT NULL,
```

```
    nome VARCHAR(25),
    categoria CATEGORIA NOT NULL,
    materiale_tubature VARCHAR(25),
    cloro BOOLEAN NOT NULL,
    anno_ultima_ristrutturazione DATE,
    caldaia VARCHAR(25),

    PRIMARY KEY (latitudine, longitudine),

    FOREIGN KEY (latitudine_stazione, longitudine_stazione) REFERENCES
Stazione_meterologica(latitudine, longitudine)
    -- gestito con trigger
);

-- Punto di prelievo
CREATE TABLE Punto_di_prelievo (
    piano INTEGER NOT NULL,
    stanza VARCHAR(15) NOT NULL,
    latitudine_sito LATITUDINE NOT NULL,
    longitudine_sito LONGITUDINE NOT NULL,
    descrizione VARCHAR(100),
    componente_idraulica VARCHAR(25) NOT NULL,

    PRIMARY KEY (latitudine_sito, longitudine_sito, piano, stanza),

    FOREIGN KEY (latitudine_sito, longitudine_sito) REFERENCES Sito(latitudine,
longitudine)
        ON DELETE RESTRICT
        ON UPDATE UPDATE
);

-- FollowUp clinico
CREATE TABLE FollowUp_clinico (
    codice CHAR(6) NOT NULL,

    PRIMARY KEY (codice)
);

-- Richiedente
CREATE TABLE Richiedente (
    codice CHAR(6) NOT NULL,
    nome VARCHAR(25),

    PRIMARY KEY (codice)
);

-- Indagine ambientale
CREATE TABLE Indagine_ambientale (
    codice CHAR(6) NOT NULL,
    codice_FollowUp CHAR(6),
    codice_Richiedente CHAR(6),
    data DATE NOT NULL,

    PRIMARY KEY (codice),

    FOREIGN KEY (codice_FollowUp) REFERENCES FollowUp_clinico(codice)
        ON DELETE SET NULL
        ON UPDATE CASCADE,
```

```
    FOREIGN KEY (codice_Richiedente) REFERENCES Richiedente(codice)
        ON DELETE SET NULL
        ON UPDATE CASCADE
);

-- Campione
CREATE TABLE Campione (
    codice CHAR(6) NOT NULL,
    longitudine_sito LONGITUDINE NOT NULL,
    latitudine_sito LATITUDINE NOT NULL,
    piano_punto_prelievo INTEGER NOT NULL,
    stanza_punto_prelievo VARCHAR(15) NOT NULL,
    codice_indagine CHAR(6) NOT NULL,
    temperatura FLOAT NOT NULL,
    matrice MATRICE NOT NULL,
    volume FLOAT_POS NOT NULL,

    PRIMARY KEY (codice),

    FOREIGN KEY (longitudine_sito, latitudine_sito, piano_punto_prelievo,
stanza_punto_prelievo) REFERENCES Punto_di_prelievo(longitudine_sito,
latitudine_sito, piano, stanza)
        ON DELETE RESTRICT
        ON UPDATE CASCADE,
    FOREIGN KEY (codice_indagine) REFERENCES Indagine_ambientale(codice)
        ON DELETE RESTRICT
        ON UPDATE CASCADE
);

-- Analisi PCR
CREATE TABLE Analisi_PCR (
    codice CHAR(6) NOT NULL,
    codice_campione CHAR(6) NOT NULL,
    data_ora DATE NOT NULL,
    esito BOOLEAN NOT NULL,
    µg_l INT_POS NOT NULL,

    PRIMARY KEY (codice),

    FOREIGN KEY (codice_campione) REFERENCES Campione(codice)
        ON DELETE CASCADE
        ON UPDATE CASCADE
);

-- Analisi colturale
CREATE TABLE Analisi_culturale (
    codice CHAR(6) NOT NULL,
    codice_campione CHAR(6) NOT NULL,
    data_ora DATE NOT NULL,
    esito BOOLEAN NOT NULL,
    ufc_l INT_POS NOT NULL,
    sierotipo VARCHAR(50),

    PRIMARY KEY (codice),

    FOREIGN KEY (codice_campione) REFERENCES Campione(codice)
        ON DELETE CASCADE
        ON UPDATE CASCADE
```

```
);

-- Analisi del pH
CREATE TABLE Analisi_pH (
    codice CHAR(6) NOT NULL,
    codice_campione CHAR(6) NOT NULL,
    data_ora DATE NOT NULL,
    ph PH NOT NULL,

    PRIMARY KEY (codice),

    FOREIGN KEY (codice_campione) REFERENCES Campione(codice)
        ON DELETE CASCADE
        ON UPDATE CASCADE
);

-- Analisi genomica
CREATE TABLE Analisi_genomica (
    codice CHAR(6) NOT NULL,
    codice_campione CHAR(6) NOT NULL,
    data_ora DATE NOT NULL,
    genoma TEXT NOT NULL,

    PRIMARY KEY (codice),

    FOREIGN KEY (codice_campione) REFERENCES Campione(codice)
        ON DELETE RESTRICT
        ON UPDATE CASCADE
);

-- Gene
CREATE TABLE Gene (
    protein_ID CHAR(6) NOT NULL,
    nome VARCHAR(75),

    PRIMARY KEY (protein_ID)
);

-- Gene del genoma
CREATE TABLE Gene_genoma (
    posizione INTEGER NOT NULL,
    codice_genoma CHAR(6) NOT NULL,
    protein_ID CHAR(6) NOT NULL,
    posizione_predecessore INTEGER,
    codice_genoma_predecessore CHAR(6),
    protein_ID_predecessore CHAR(6),
    query_cover PERCENT NOT NULL,
    percent_identity PERCENT NOT NULL,
    e_value FLOAT_POS NOT NULL,

    PRIMARY KEY (posizione, codice_genoma, protein_ID),

    FOREIGN KEY (codice_genoma) REFERENCES Analisi_genomica(codice)
        ON DELETE CASCADE
        ON UPDATE CASCADE,
    FOREIGN KEY (protein_ID) REFERENCES Gene(protein_ID)
        ON DELETE RESTRICT
        ON UPDATE CASCADE,
```

```
    FOREIGN KEY (posizione_predecessore, codice_genoma_predecessore,
protein_ID_predecessore) REFERENCES Gene_genoma(posizione, codice_genoma,
protein_ID)
        ON DELETE SET NULL
        ON UPDATE CASCADE
);
```

7. Bibliografia

[1] M. della Salute, «Linee guida per la prevenzione e il controllo della legionellosi», 2015, [Online]. Disponibile su: https://www.salute.gov.it/portale/malattieInfettive/dettaglioPubblicazioniMalattieInfettive.jsp?id=2362

[2] A. Felice, M. Franchi, S. De Martin, N. Vitacolonna, L. Iacumin, e M. Civilini, «Environmental surveillance and spatio-temporal analysis of Legionella spp. in a region of northeastern Italy (2002–2017)», *PLOS ONE*, vol. 14, fasc. 7, p. e218687, 2019, doi: 10.1371/journal.pone.0218687.

[3] D. Garlatti, «Base di dati e applicazione web per il monitoraggio del batterio della legionella», 2020.

[4] P. Atzeni, S. Ceri, S. Paraboschi, e R. Torlone, *Database Systems: Concepts, Languages & Architectures*. 1999.

[5] A. Colautti *et al.*, «Draft genome sequences from 127 Legionella spp. strains isolated in water systems linked to legionellosis outbreaks», *Microbiol Resour Announc*, vol. 13, fasc. 6, 2024, doi: 10.1128/mra.01154-23.

[6] «Polymerase chain Reaction (PCR)». [Online]. Disponibile su: https://www.genome.gov/genetics-glossary/Polymerase-Chain-Reaction

8. Glossario

Al fine di facilitare la comprensione del documento, è redatto il seguente glossario contenente le definizioni dei termini tecnici utilizzati.

Termine	**Definizione**
Aerosol	Particelle sospese nell'aria, contenenti gocce d'acqua, che possono trasportare il batterio Legionella.
Analisi	Esame di laboratorio effettuato su campioni di acqua prelevati durante un'indagine ambientale.
Analisi Colturale	Esame di laboratorio che permette di isolare e identificare le unità formanti colonia (UFC_L) di Legionella in un campione di acqua.
Attributo	Concetto che descrive una proprietà o una componente di una entità o di una relazione. (*i.e.* campo).
Attributo composto	Attributo dalla struttura complessa, costituito da diversi sotto-attributi.
Attributo multivalore	Attributo che, per ogni istanza dell'entità cui è associato, può assumere più di un valore.
Campione	Piccola quantità di acqua da sottoporre a esame.
Categoria	Classificazione di un sito, o più specificamente di un edificio, in base alla sua destinazione d'uso, come ad esempio ospedaliero, termale o alberghiero.
Chiave primaria	Attributo o insieme di attributi che identifica univocamente ogni istanza di un'entità.

Glossario

Termine	Definizione
Componente idraulica	Componente di un sistema idraulico da cui viene prelevato un campione di acqua, come un rubinetto o un filtro di un impianto di condizionamento.
Entità	In riferimento allo schema E-R, descrive una classe di oggetti con esistenza autonoma, con particolare significato nel contesto in esame. (*i.e.* tabella).
Entità debole	Entità che non ha una chiave primaria propria, ma dipende da un'altra entità per la sua identificazione.
Generalizzazione	In riferimento al modello E-R, relazione che associa ad un'entità genitore una o più entità figlie, che ereditano le proprietà del genitore. (*i.e.* specializzazione).
FollowUp Clinico	Indagine ambientale, o indagini ambientali, condotte a seguito di uno o più casi di legionellosi. Tali indagini non si limitano al domicilio del paziente, ma possono estendersi a tutti i luoghi frequentati dal malato nei dieci giorni precedenti l'insorgenza dei sintomi. La decisione di effettuare tali indagini è lasciata al competente servizio territoriale che «deve valutare di volta in volta l'opportunità di effettuare o meno dei campionamenti ambientali, sulla base della valutazione dei rischio»[22].
Indagine Ambientale	Collezione di campioni prelevati da un sito specifico in una data specifica.

Glossario

[22] [1], «Linee guida per la prevenzione ed il controllo della legionellosi», p. 30

Termine	Definizione
PCR	Polymerase Chain Reaction, è una «tecnica di laboratorio per produrre rapidamente (amplificare) milioni o miliardi di copie di uno specifico segmento di DNA, che può poi essere studiato in modo più dettagliato. La PCR prevede l'uso di brevi frammenti di DNA sintetico chiamati primer per selezionare un segmento del genoma da amplificare, e quindi più cicli di sintesi del DNA per amplificare quel segmento»[23].
PCR Qualitativa	Esame di laboratorio che fornisce un'informazione dicotomica sulla presenza di Legionella in un campione.
PCR Quantitativa	Esame di laboratorio rapido che rileva e quantifica il DNA o l'RNA di Legionella presenti in un campione. (*i.e.* Real-Time PCR o qPCR).
Relazione	In riferimento allo schema E-R, legame che rappresenta la connessione logica e significativa per la realtà modellata, tra due o più entità.
Relazione Ricorsiva	Relazione che associa una entità a se stessa (*i.e.* relazione autoreferenziale).
Richiedente	Ente o istituzione che richiede un'indagine ambientale.
Sierotipo	Livello di classificazione di batteri di Legionella inferiore a quello specie. Il laboratorio ARPA distingue tre sierotipi: sierotipo 1, sierotipo 2-15 e sierotipo sp (*i.e.* sierogruppo).

Glossario

[23][6] «Polymerase chain Reaction (PCR)». [Online]. Disponibile su: https://www.genome.gov/genetics-glossary/Polymerase-Chain-Reaction

Termine	Definizione
Sito	Edificio presso il quale è condotta un'indagine ambientale.
UFC_L	Unità formanti colonie per litro: ovvero unità di misura utilizzata per indicare la concentrazione di Legionella in un campione d'acqua destinato all'analisi colturale.
UG_L	Microgrammi per litro: ovvero unità di misura utilizzata per determinare la concentrazione di Legionella in un campione d'acqua mediante PCR quantitativa.

Glossario