Analisi dei flussi e mappatura delle banche dati di interesse per la task force dati per l'emergenza COVID-19

a cura del Gruppo di lavoro 2 - Data collection and Infrastructure

Sommario

1. Introduzione

Il Gruppo di lavoro 2 ha come obiettivo principale quello di mappare, raccogliere e ottimizzare i flussi di gestione di una serie di dataset/banche dati di interesse che possono essere utilizzati nell'emergenza per abilitare analisi avanzate e predittive, utili per supportare i decisori pubblici nelle scelte strategiche.

Sono tre i macro-obiettivi che il Gruppo di lavoro 2 è chiamato a raggiungere:

1. Analisi e mappatura delle banche dati di interesse e dei livelli di interoperabilità;
2. Raccolta e consolidamento dei dati necessari allo svolgimento delle attività dei sottogruppi;
3. Messa a sistema e/o ottimizzazione dei flussi e dei tempi dei dati sanitari e amministrativi regionali; gestione della interoperabilità, per la realizzazione di un ecosistema basato su tecnologie di big data

Questo documento ha l'obiettivo di descrivere le attività svolte, illustrando la metodologia impiegata ed evidenziando, con particolare riguardo al contesto sanitario, i primi risultati di un'analisi condotta su alcuni dataset specifici.
Per agevolare la lettura del documento, si fornisce una sintetica descrizione della sua struttura:

- la sezione 2 illustra la metodologia di mappatura dei dati impiegata;
- la sezione 3 descrive i domini verticali individuati;
- la sezione 4 si concentra sul dominio sanitario presentando una prima analisi di alcuni dataset individuati come prioritari per rispondere a fabbisogni specificamente individuati.

2. Metodologia mappatura dati

La metodologia impiegata dal Gruppo di lavoro 2 si è contraddistinta con una fase di raccolta delle richieste/bisogni/domande espresse dagli altri sottogruppi; a seguito di questa, è stato poi svolto un lavoro per individuare il corretto interlocutore di competenza a cui formulare la richiesta. Questo ha comportato una breve analisi, attraverso i siti web istituzionali di pubbliche amministrazioni e di enti privati, dei dati già pubblicati o di competenza rispetto anche a specifici riferimenti legislativi. Una volta individuato l'ente (sia esso privato che pubblico), è stata predisposta una lettera ufficiale a firma Ministro per l'Innovazione e Digitalizzazione (MID) inviata via canali PEC. Ricevuti i dati dagli enti contattati,

il Gruppo di lavoro 2 ha provveduto al trasferimento degli stessi, previa analisi o rielaborazione in alcuni casi, al sottogruppo richiedente.
Contestualmente, il Gruppo di lavoro 2, essendo composto da attori strategici nella catena decisionale sull'emergenza COVID-19 (i.e., Ministero della Salute e Dipartimento di Protezione Civile) ha evidenziato tre scenari di analisi strategiche che sistematizzate consentirebbero di formulare una risposta in tempi adeguati alla gestione di una situazione di emergenza.

La Figura 1 illustra i passi della metodologia finora applicata per la mappatura dei dati.

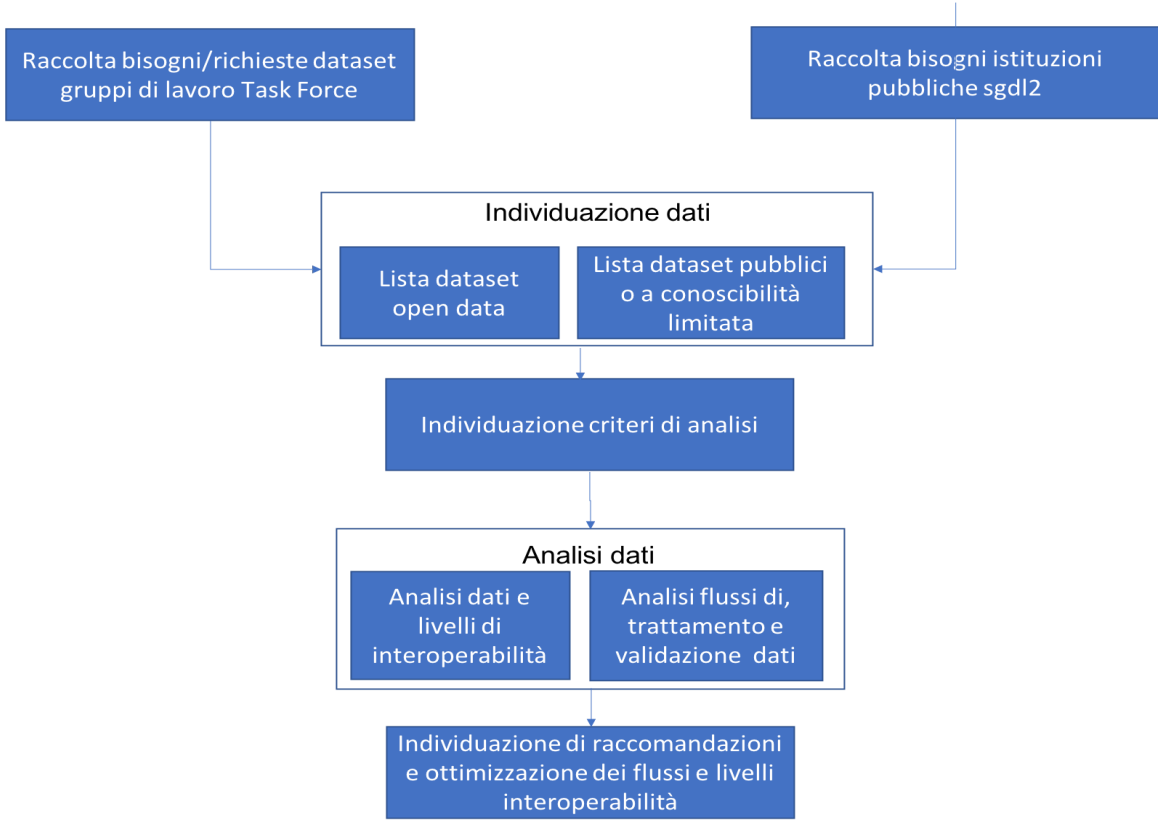


Figura 1: Metodologia mappatura dati

L'analisi, condotta su alcuni dataset in ambito sanitario, ha evidenziato i seguenti flussi di trattamento e validazione dei dati:

1. Flussi (**individuali e aggregati)** già esistenti nel patrimonio informativo Nuovo Sistema Informativo Sanitario (NSIS), (secondo la mappatura messa a disposizione dal Ministero della Salute) e flussi ad hoc attivati per la gestione emergenziale al fine di affrontare le esigenze immediate sia della fase 1 che della fase 2.
   I flussi attivati ad hoc, rispetto a quelli tradizionali, richiedono una maggiore tempestività di raccolta (frequenza di aggiornamento del dato *real time*, giornaliera, bisettimanale o al massimo settimanale), sembrano diretti a intercettare tipologie di dati per lo più provenienti dai territori, e sono ulteriori flussi paralleli rispetto a quelli tradizionali.

2. Flussi dati interconnessi **strategici**: per poter affrontare azioni nel medio e lungo periodo; risulterebbe quantomai necessario, anche sulla scorta dell'esperienza acquisita, la predisposizione di un aggiornamento dei sistemi esistenti a eventuali future emergenze analoghe. In tal caso i flussi tradizionali, migliorati e arricchiti con le informazioni emergenziali, con le raccomandazioni formulate unitamente allo storico dei dati, possono rappresentare per le Amministrazioni un supporto di grande valore.

3. Domini individuati

La Figura 2 illustra i domini verticali attualmente in fase di analisi da parte del Gruppo di lavoro 2. In particolare, sono individuati i domini per cui sono stati già raccolti alcuni dataset e individuati specifici flussi provenienti da alcuni sistemi esistenti. (e.g., in quest'ultimo caso dal Nuovo Sistema Informativo Sanitario del Ministero della Salute).

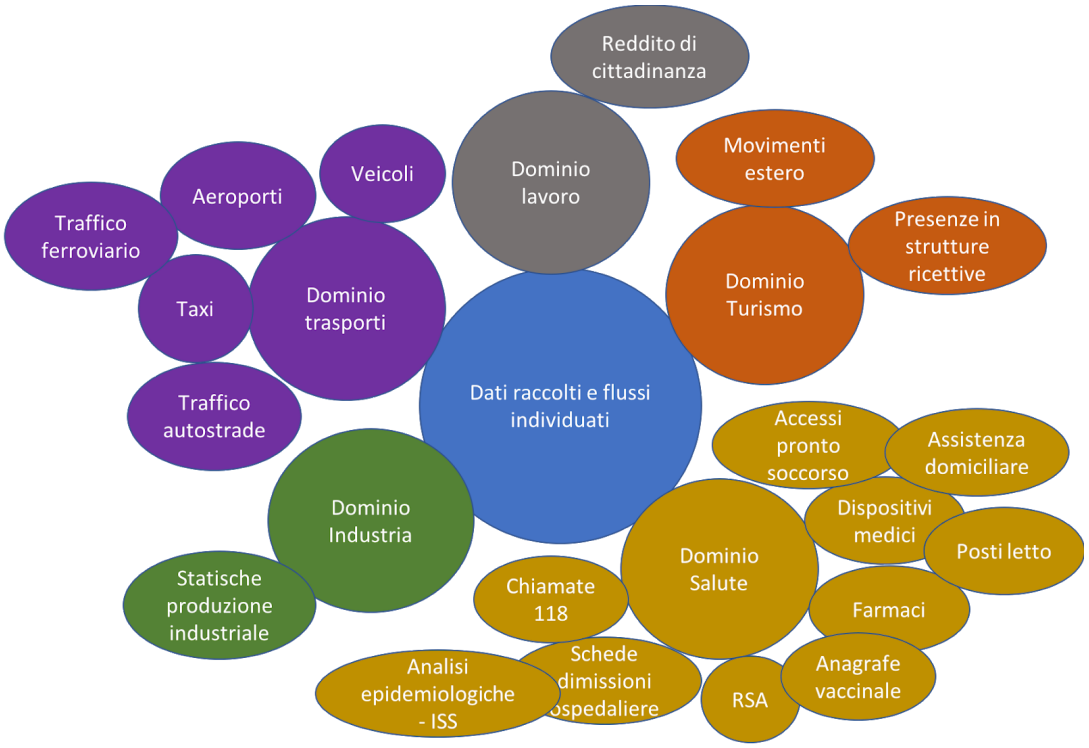


Figura 2: Dati raccolti e flussi individuati

Con riferimento ai flussi sanitari trasmessi dal Ministero della Salute, il Gruppo di lavoro 2 ne ha mappati solo alcuni di tipo aperto, alcuni dei quali recentemente aggiornati, e ha evidenziato in casi specifici criticità per la condivisione dei dati provenienti dai flussi NSIS. Anche in relazione alle problematiche di protezione dei dati personali, appare auspicabile una collaborazione più stretta con il Gruppo di lavoro 8 "Profili giuridici della gestione dei dati connessa all'emergenza" e la definizione di raccomandazioni anche per eventuali interventi normativi, comunque in atto presso il Ministero della Salute.

3.1 Interazione con i sottogruppi della task force

Le prime richieste di raccolta dati sono state fatte pervenire dal Gruppo di lavoro 3 “Impatto economico”, responsabile delle analisi di impatto economico. Le richieste riguardavano una lista di dataset che spaziavano diversi domini applicativi, dal dominio dei trasporti a quello del turismo, del lavoro e dell’industria nonché del dominio energia con i relativi consumi. Con le medesime modalità già illustrate, mediante richieste formali da parte del Ministro per l’Innovazione e la Digitalizzazione (MID) indirizzate alle istituzioni competenti per ciascun dataset, sono stati ricevuti sia i contatti dei referenti tecnici che i dataset da analizzare. Come si nota dalla Figura 2 il dominio dei trasporti è quello con più dataset finora ricevuti. Per tali dataset il Gruppo di lavoro 2 ha avviato una serie di analisi mirate sia ad ottimizzazione i file ricevuti, sia a documentare le loro caratteristiche sulla base dei seguenti criteri di analisi:

- copertura temporale come richiesta;
- copertura geografica come richiesta;
- caratteristiche dei campi del dato (e.g., stringhe, uso di date, quante colonne sono presenti, uso di vocabolari controllati, uso di codici univoci);
- formato dei dataset ed encoding;
- qualità dei dati in termini principalmente di accuratezza sintattica (e.g., presenza di “null”, di righe vuote, altri elementi sporchi, uso di maiuscole e minuscole, accenti/apostrofi), semantica (e.g., è chiara la semantica delle colonne, si usa un valore che non è in linea con la semantica suggerita dall’intestazione delle colonne, uso di vocabolari standard nazionali, ecc.).

Per tutti questi domini, è stata avviata anche una ricerca di open data partendo dal paniere dinamico di dataset pubblicato dall’Agenzia per l’Italia Digitale per ciascuna regione, e verificando la disponibilità di alcuni dataset di singole amministrazioni (come nel caso di ISTAT o del MEF).

Si sottolinea che, il Ministro per la l’Innovazione e la Digitalizzazione (MID) e il suo staff sono attualmente in contatto anche con ARERA (Autorità di Regolazione per Energia Reti e Ambiente) per la raccolta di dati sul consumo energetico di interesse per il Gruppo di lavoro 3 “Impatto economico”. I dati offerti sono aggregati per alta, media e bassa tensione e non per specifico settore di attività ATECO.

Successivamente, il Gruppo di lavoro 7 "Big data & AI for policies" ha fatto pervenire il bisogno di accedere ad alcuni dati sanitari in possesso dell'Istituto Superiore di Sanità (ISS) per un'analisi epidemiologica che possa consentire di comprendere l'impatto della diffusione del COVID-19. I dati richiesti riguardavano:

- casi di COVID-19 totali rilevati mediante i campioni processati dal Laboratorio nazionale di riferimento presso l'istituto;
- casi di COVID-19 relativi a operatori sanitari;
- numero dei deceduti suddivisi per fascia di età con relativo calcolo della letalità;

e sono aggregati per:

- età
- sesso
- luogo di residenza
- tempo intercorso (in giorni) tra la data di effettuazione del test e l'esito
- tempo di ricovero in giorni e dopo quanto in media si manifestano i sintomi più gravi ospedale e reparto
- patologie preesistenti: quali e quante suddivise per età e sesso
- diagnosi di ricovero e decorso clinico con specificazione del trattamento ricevuto (trasferimento in terapia intensiva, uso di ossigeno e ventilazione) e delle eventuali complicanze

Infine, grazie a interazioni con il Gruppo di lavoro 4 "Web data e impatto socio-economico" in merito a possibili analisi di impatto socio-economico, anche in relazione a uno degli scenari di analisi predisposti dal Gruppo di lavoro 2 su pazienti fragili ai tempi del COVID-19 (si veda la sezione 4.2), sono pervenute richieste di accesso ai dati sul reddito di cittadinanza, utili per comprendere il livello di fragilità economica dei cittadini, ponendoli in relazione ad altri dati sulla mobilità o sull'uso di certe tipologie di farmaci, per citare un paio di esempi.

Per concludere, oltre ai domini attualmente mappati in base a specifiche richieste di altri gruppi, il Gruppo di lavoro 2 ha discusso la possibilità di lavorare su altri domini che possono presentare interessanti correlazioni con i dati raccolti dal Ministero della Salute e dal Dipartimento di Protezione Civile. In particolare, si è discusso del:

- **Dominio Ambientale**: con i dati del Sinanet dell'ISPRA e di Copernicus per verificare quelli relativi alla qualità dell'aria e al meteo.
- **Dominio Statistico/Demografico** con i dati sul censimento della popolazione residente anche straniera
- **Dominio Protezione Civile** con i dati sull'acquisto dei materiali (i cui contratti sono stati pubblicati in open data dal Dipartimento della Protezione Civile), sul triage Campale, sulle strutture di quarantena, sugli ospedali campali.

3.2 Monitoraggio richieste dati

Nell'ambito dell'attività di raccolta e consolidamento dei dati necessari allo svolgimento delle attività dei sottogruppi sono stati contattati 15 soggetti, 3 dei quali aziende private. Per quanto riguarda i sottogruppi della task force che hanno richiesto i dati sui quali sono state poi formalizzate le richieste, la maggior parte sono pervenute dal Gruppo di lavoro 3 "Impatto economico" e dal Gruppo di lavoro 7 "Big data & AI for policies"; in particolare da quest'ultimo proviene la richiesta dei dati aperti in possesso dell'ISS.

La maggior parte di essi è stata contattata in data 21 marzo 2020 tramite comunicazione inviata dall'ufficio di gabinetto del ministro.

Nel dettaglio i soggetti contattati sono i seguenti:

Enti/istituzioni

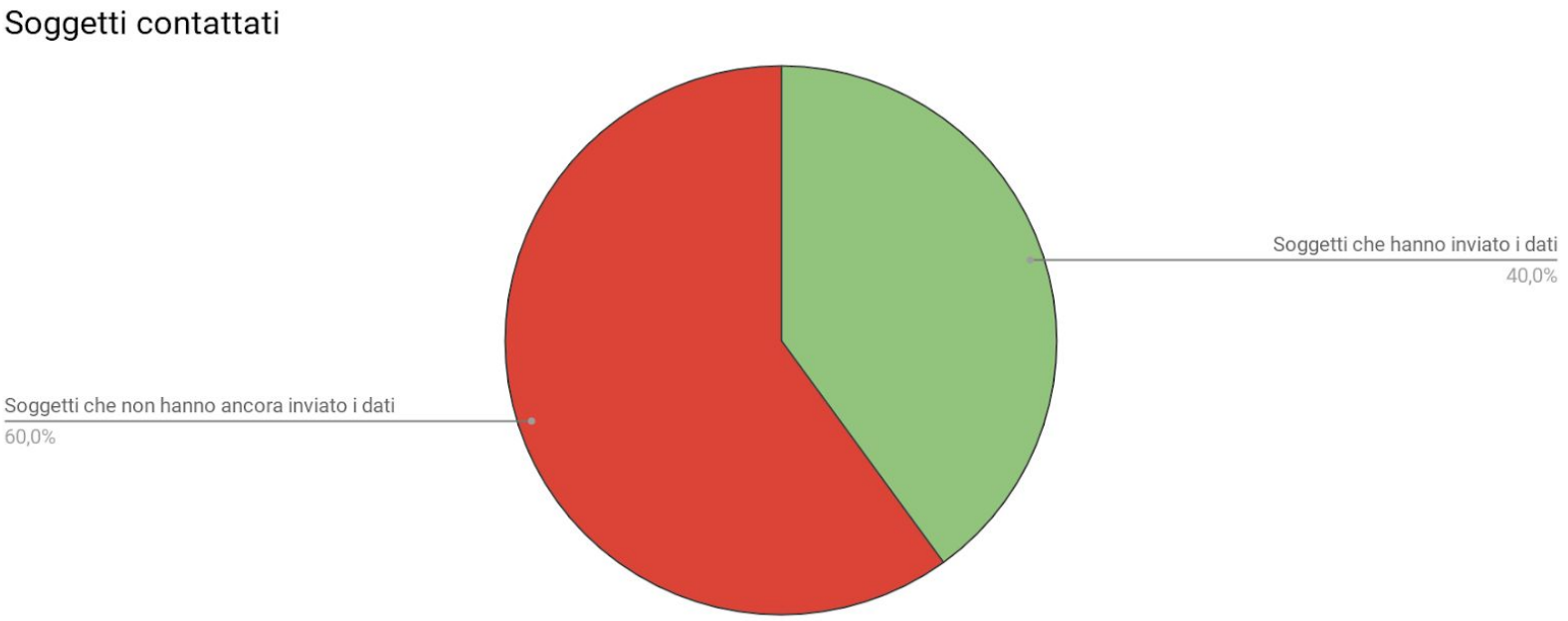
- AREA
- ASSAEROPORTI
- ACI
- Agenzia Entrate
- ENIT
- INPS
- IRPET
- ISS
- ISTAT

- MISE
- MIT
- Ministero lavoro e politiche sociali

Operatori privati:
- CRIF S.p.A.
- TRENITALIA
- ITALO

In data 29 maggio 2020, sono 11 i soggetti che hanno risposto alla richiesta inviata dal Ministro Pisano esprimendo la disponibilità a collaborare.

A fronte della risposta, i soggetti che successivamente hanno fornito in parte o completamente i dati richiesti (seppure in alcuni casi con qualche differenza dalla richiesta iniziale) sono 6 su 15.



Nei casi di mancata risposta alla richiesta iniziale, è stato inviato un sollecito tramite l'Ufficio di Gabinetto del Ministro per l'innovazione tecnologica e la digitalizzazione.

Con alcune pubbliche amministrazioni, come INPS e ARERA, ci sono stati scambi recenti e gli enti stanno procedendo all'estrazione dei dati richiesti. Con il MIT, è stata avviata un'interlocuzione dalla fine di aprile, che ha portato all'individuazione di referenti per i vari gruppi e condivisi gli obiettivi della Task Force con l'obiettivo di condividere una lista di temi e problematiche da affrontare insieme trasversalmente con i gruppi della task force. Ad oggi non ci sono ancora proposte operative o richieste di supporto. In questo quadro, altri soggetti si sono resi disponibili a supportarci, oltre a fornirci spunti di analisi e aggregati dati sul tema trasporti, come nel caso dell'Autorità dei Trasporti.
Infine, è stato recentemente avviato con CRIF S.p.A. un percorso che vedrà, già nei giorni successivi al rilascio di questo primo report, un incontro per stabilire meglio i dettagli delle richieste avanzate e condivise.

4. Dominio sanitario

Nel contesto del dominio sanitario i lavori si sono articolati in due fasi:

4.1 Fase A

Questa fase si è connotata per l'attività di individuazione dei flussi di dati provenienti dal Nuovo Sistema Informativo Sanitario (NSIS). In particolare, i flussi elencati dal Gruppo di lavoro 2 e poi successivamente catalogati con diversi livelli di priorità sono i seguenti:

- **priorità A:**
  - Scheda Dimissione Ospedaliera (SDO) – frequenza di aggiornamento mensile
  - Consumo dei medicinali in ambito ospedaliero nelle strutture pubbliche del SSN (MEDOSP) – frequenza di aggiornamento mensile
  - Tracciabilità del farmaco – produzione e distribuzione intermedia (TRACCIA) – frequenza di aggiornamento giornaliera e a evento
  - Dispositivi medici – consumi nelle strutture pubbliche del SSN (DISPMED) - frequenza di aggiornamento mensile e a evento

- **priorità B:**
  - Accessi al pronto soccorso (EMUR-PS) – frequenza di aggiornamento mensile
  - Eventi sanitari del 118 (EMUR-118) – frequenza di aggiornamento mensile
  - Accessi sanitari al domicilio dell'assistito (SIAD) – frequenza di aggiornamento mensile/trimestrale

- **priorità C:**
  - Grandi apparecchiature (GAP) – frequenza di aggiornamento a evento. Nel contesto delle grandi apparecchiature esistono poi i flussi HSP14 e STS14 che riguardano specificatamente apparecchiature di pertinenza per il COVID-19.

- **priorità D:**
    - Anagrafe vaccinale nazionale (AVN) – frequenza di aggiornamento trimestrale (questo dataset è stato ritenuto importante ma non in questa fase, ma più avanti quando un vaccino sul COVID-19 sarà disponibile).

Oltre ai flussi sopra evidenziati ne sono stati elencati altri del sistema NSIS ma ritenuti non necessari in questa fase A. Questi sono per esempio:

i) Dipendenze (assistenza rivolta a persone dipendenti da sostanze stupefacenti o alcool o psicotrope;
ii) Salute mentale (interventi sanitari a persone adulte con problemi psichiatrici);
iii) Hospice (assistenza sanitaria e socio-sanitaria in hospice).

Nel contesto dei dati del Dipartimento di Protezione Civile, oltre ai già noti dati aperti pubblicati in repository Github e aggiornati ogni giorno per fornire le informazioni del bollettino pomeridiano delle 18:00, il Dipartimento ha avviato una collaborazione, non ancora formalizzata al momento del rilascio di questo documento, con l'Istituto Superiore di Sanità (ISS). In tale collaborazione alcuni dati dell'ISS dovrebbero essere aperti attraverso il dipartimento di protezione civile. A tal proposito un primo tracciato record è stato proposto nel contesto di questa collaborazione tra i due enti.

4.2 Fase B

Nella Fase B, il Ministero della Salute e il Dipartimento di Protezione Civile hanno identificato una serie di scenari e bisogni di analisi da effettuare sui dati. Tali scenari possono essere riassunti come segue:

- analisi dei dati di struttura per verificare la trasformazione dell'offerta ospedaliera a seguito dell'emergenza COVID-19;
- analisi dei farmaci utilizzati dai pazienti no COVID-19 durante la pandemia e confronto con i consumi del primo trimestre dello scorso anno (per i diversi canali di distribuzione del farmaco);
- analisi dell'impatto socio-sanitario del COVID-19 sui pazienti fragili (per età, sesso, cronicità, nazionalità, titolo di studio - ove disponibile) in termini di prescrizioni farmacologiche, prestazioni ambulatoriali, accessi di pronto soccorso, rispetto al primo trimestre dello scorso anno e per ambito territoriale (Comune/Provincia);
- strutturazione di una lettura a livello geografico dei diversi dati (reddituali, demografici, sanitari) prendendo a riferimento le diverse fonti dati istituzionali disponibili.

A tal proposito, il Gruppo di lavoro 2 ha voluto riassumere in una visione complessiva i diversi elementi che sono emersi nella fase A e B. Pertanto, per ciascuno degli scenari di analisi suddetti (fase B) sono stati mappati i flussi del sistema NSIS e altri individuati (fase A), e per ciascuno di questi, sono stati indicati dei criteri che possono contribuire a prioritizzare le analisi da svolgere anche in collaborazione con altri gruppi della task force, già contattati a tale scopo (e.g., nel caso dell'analisi socio-sanitaria si è stabilito un contatto con il Gruppo di lavoro 4 "Web data e impatto socio-economico").
Per ciascun flusso coinvolto negli scenari di analisi si è visto quindi se:

- dati aperti sono disponibili e se sono aggiornati;
- dati esterni sono richiesti e se sono dati aperti aggiornati (esempi di dati esterni sono: le statistiche sulla popolazione residente, le statistiche del livello di istruzione, i dati sulle dichiarazioni dei redditi);
- dati personali sono coinvolti nel flusso;
- esistono specifici vincoli temporali (magari imposti anche dalla disponibilità di qualche norma che disciplina la possibile messa a disposizione dei flussi per elaborazioni).

Infine, un'importante opportunità nella lettura dei dati dell'emergenza potrebbe essere rappresentata dall'utilizzo del modello nazionale di stratificazione della popolazione, in fase di definizione presso il Ministero della Salute. Attraverso l'interconnessione dei dati individuali si potranno costruire scenari tendenziali che tengano conto degli attuali profili di consumo della popolazione suddivisa per patologie croniche e verificare l'impatto generato da una situazione pandemica sul sistema e sui diversi profili di consumo della popolazione cronica.
A tal fine, il Ministero della Salute sta sviluppando un lavoro articolato sul tema dell'utilizzo innovativo dei dati sanitari, in linea con le linee programmatiche dell'UE in materia di digitalizzazione della sanità e analisi dei Big Data. In questo contesto, il Ministero ha presentato al Gruppo di lavoro 2 il "Modello Predittivo", un progetto che ha l'obiettivo di costruire nel 2020 uno scenario programmatico sul fabbisogno di salute della popolazione italiana e della spesa sanitaria, simulando scenari predittivi a medio-lungo termine, partendo dalle tendenze in atto. Questo progetto, basato proprio sul modello di stratificazione della popolazione suddetto, prevede anche un "Modulo Covid-19" dedicato alla pandemia in corso. Il Ministero della Salute sta implementando anche un progetto dedicato agli IRCCS (Istituti di Ricerca e Cura a Carattere Scientifico) denominato "Health Big Data" che prevede la realizzazione di una infrastruttura IT centralizzata che consenta la gestione dei dati generati dal progetto e la messa in rete di tutti gli IRCCS. Il progetto prevede, inoltre, la realizzazione di una unità di Data Infrastructure e una di Data Analysis.

Nell'ambito delle strategie per la digitalizzazione e l'utilizzo innovativo dei dati sanitari un ruolo fondamentale è riservato allo sviluppo e al radicamento sul territorio nazionale del Fascicolo Sanitario Elettronico (FSE), anche al fine di dare la possibilità ai cittadini di tenere costantemente sotto controllo i propri dati sanitari, arricchendo ulteriormente il patrimonio dei dati sanitari del Ministero della Salute.

4.3 Infrastrutture

Nell'ambito delle attività del GdI2 sono state analizzate principalmente due infrastrutture: il sistema NSIS - Nuovo Sistema Informativo Sanitario di titolarità del Ministero della Salute, e l'infrastruttura dati che la Protezione Civile ha messo in campo per la pubblicazione e comunicazione dei dati giornalieri sull'andamento nazionale dell'epidemia COVID-19.

4.3.1 Infrastruttura NSIS

Il sistema NSIS è la più importante banca dati sanitaria a livello nazionale a supporto della programmazione sanitaria nazionale e regionale. I dati del NSIS, una volta sottoposti a controlli di completezza, tempestività e qualità consentono la definizione di vari modelli previsionali e indicatori finalizzati all'interpretazione di specifici fenomeni sanitari, nonché analisi quantitative orientate alla simulazione dell'effetto di interventi normativi tra i quali, a titolo esemplificativo, la revisione della normativa sulla compartecipazione alla spesa e di quella relativa alle tariffe associate al nomenclatore di specialistica ambulatoriale.
Inoltre, il patrimonio dei dati dell'NSIS è utilizzato per monitorare:

- l'assistenza relativa all'emergenza sanitaria;
- gli "eventi sentinella" che si verificano nelle strutture del Sistema Sanitario Nazionale (SSN), al fine di comprendere le circostanze e i fattori che ne hanno favorito l'occorrenza.
- le liste di attesa che rappresentano in tutti i sistemi sanitari uno dei problemi maggiormente avvertiti dai cittadini

In sostanza, il NSIS raccoglie in maniera omogenea i dati prodotti a livello regionale e locale relativi alla domanda e all'offerta di assistenza sanitaria consentendo alle diverse componenti del SSN di dialogare tra loro (interoperabilità). Il livello regionale e locale conferisce i dati in proprio possesso mediante caricamento di file XML presso la piattaforma NSIS, piattaforma attualmente ospitata su cloud di INAIL. Il caricamento dei file avviene previa autenticazione presso il sistema

NSIS di utenti autorizzati e consiste in un'operazione di upload di file XML, validati successivamente dal sistema rispetto agli schemi XSD dei vari flussi.
Grazie alla standardizzazione operata sui flussi del sistema, il Ministero della Salute ha provveduto anche al rilascio di alcuni dei dati secondo il paradigma degli open data. Nello specifico, il Ministero mette a disposizione dal 2012 un'infrastruttura open source già realizzata nel contesto di Open Lab di Microsoft Corp. L'infrastruttura utilizza Microsoft Azure per la memorizzazione dei dati che possono essere interrogati mediante API SOAP-REST. Al momento, i dataset aperti del Ministero non risultano essere metadatati secondo il profilo nazionale di metadatazione DCAT-AP_IT e quindi non presenti nel catalogo nazionale dei dati aperti dati.gov.it.

4.3.2 Infrastruttura dati Protezione Civile

Il Dipartimento di Protezione Civile ha deciso di adottare la seguente metodologia di pubblicazione dei dati sul COVID-19 da essa raccolti, rendendosi quasi un caso unico in Italia:

- pubblicazione dei dati aperti con licenza CC-BY 4.0 su piattaforma Github, suddivisi per livello regionale e provinciale. La piattaforma consente un'apertura verso gli utenti finali che possono aprire *issue* per interagire con i responsabili dei dati in merito alla pubblicazione. I dati pubblicati sono sia quelli del bollettino giornaliero, sia i dati sui contratti di forniture di materiali inerenti la pandemia, sia i dati relativi alle zone dichiarate rosse (in questo caso shapefile);
- pubblicazione di una breve descrizione per ciascun dato del tracciato record definito;
- pubblicazione dei metadati secondo gli standard RNDT/INSPIRE per i dati geospaziali e DCAT-AP_IT per gli altri tipi di dati su piattaforma Github, in una directory specifica per la metadatazione. Questo ha consentito al Dipartimento di Protezione Civile di interagire mediante procedure di harvesting automatico con i due cataloghi nazionali dei dati, dati.gov.it per i dati aperti e geodati.gov.it per i dati di tipo geospaziale. I dataset quindi da loro pubblicati sono regolarmente documentati nei cataloghi nazionali, grazie proprio all'implementazione dei profili standard di metadatazione che consentono un interscambio interoperabile dei metadati;

- pubblicazione di mappe e infografiche che visualizzano facilmente l'andamento del contagio COVID-19 e che utilizzano i dati *raw* pubblicati su Github. In questo caso, è stata utilizzata l'infrastruttura arcgis[1] disponibile nella versione desktop e nella versione per dispositivi mobili.

4.4 Analisi flussi dati COVID-19 provenienti dal livello locale

La sorveglianza comprende tutti i casi di COVID-19 diagnosticati dai laboratori di riferimento regionali, come indicato dalla Circolare del Ministero della Salute n. 0005443 del 22 febbraio 2020. I dati individuali, aventi un livello di dettaglio maggiore rispetto a quello previsto in altri flussi esistenti al momento, vengono aggiornati da ciascuna Regioni/PPAA con cadenza giornaliera.

È da evidenziare che il completamento delle informazioni a livello centrale (Ministero della Salute, alla Protezione Civile ed all'Istituto Superiore di Sanità) che vengono raccolte e trasmesse da parte delle Regioni/PPAA può richiedere qualche giorno, viste le modalità diverse di trasmissione degli stessi e dei relativi flussi. Non deve quindi sorprendere la discordanza con quanto riportato attraverso altri flussi informativi che raccolgono dati aggregati con minor livello di dettaglio. I dati raccolti sono in continua fase di consolidamento e, come prevedibile in una situazione emergenziale, alcune informazioni acquisite specialmente dall'Istituto Superiore di Sanità (ISS), sono incomplete.

4.4.1 Flusso informatico a livello nazionale - NOTIFICA DELLE MALATTIE INFETTIVE

Il Servizio definito all'interno di ogni Regione (Sistema Sanitario Regionale) è impegnato nel debito informativo quotidiano sui dati della sorveglianza COVID-19 al Ministero della Salute, alla Protezione Civile ed all'Istituto Superiore di Sanità. Il flusso ufficiale dei dati segue tre direttrici:

- comunicazione giornaliera al Ministero della Salute, Protezione Civile nazionale, al massimo entro le ore 15,00 dei dati numerici relativi ai casi, tamponi effettuati, ecc., …
  Il tracciato è definito dalla protezione civile ed è quello documentato nella piattaforma Github prima descritta.
- secondo quanto stabilito dal Regolamento Sanitario Internazionale, le Aziende Sanitarie Provinciali/Aziende Ospedaliere territorialmente competenti inviano una scheda di notifica di tutti i casi che corrispondono alla

[1] http://opendatadpc.maps.arcgis.com/apps/opsdashboard/index.html#/b0c68bce2cce478eaac82fe38d4138b1

definizione di caso entro 24 ore dalla rilevazione, la Regione provvede al suo invio al Ministero della Salute - Direzione Generale della Prevenzione sanitaria (Ufficio 5 - Prevenzione delle Malattie Trasmissibili e Profilassi Internazionale) e all'Istituto Superiore di Sanità (Dipartimento di Malattie Infettive), previa registrazione sul sito web dedicato. Oltre alle informazioni contenute nella scheda di notifica, devono essere raccolte anche informazioni, che permettano l'attivazione di tutte le misure di sanità pubblica;

- secondo l'Ordinanza della Protezione Civile n. 640 del 27 febbraio 2020, che recita al punto 3. "*È fatto obbligo alle Regioni/PPAA di alimentare quotidianamente la piattaforma dati (ISS) di cui al comma 2, caricando entro le ore 11.00 di ogni giorno i dati relativi al giorno precedente*", i dati riferiti all'indagine epidemiologica (comune, luogo di esposizione, eventuali casi collegati, se il caso è un operatore sanitario), data e laboratorio di esecuzione del tampone rinofaringeo, caratteristiche cliniche e condizioni predisponenti, stato del paziente e luogo di ricovero (ospedale, domicilio) ed allineandoli con i dati che la Protezione Civile comunica giornalmente prima tramite conferenza stampa ora tramite il rilascio dei dati e dashboard alla stessa ora.

In particolare, gli artt. 1 e 2 della predetta Ordinanza rispettivamente rubricati "Sorveglianza epidemiologica" e "Sorveglianza microbiologica" disciplinano i compiti affidati all'Istituto Superiore di Sanità nonché le modalità di raccolta e trasmissione dei dati da parte delle Regioni e delle Province autonome di Trento e Bolzano. Più precisamente, l'art. 1 prevede una specifica piattaforma gestita dall'Istituto Superiore di Sanità, alimentata dai dati trasmessi dalle Regioni e dalle Province autonome quotidianamente; l'art. 2 definisce le modalità di raccolta dei campioni biologici delle persone sottoposte ad indagine epidemiologica e le modalità di verifica e conservazione dei dati da parte dell'Istituto Superiore di Sanità.
Il combinato disposto dell'art. 1 e del successivo art. 3 consente inoltre di individuare nell'Istituto nazionale di malattie Infettive Lazzaro Spallanzani di Roma, il soggetto che in collaborazione con l'Istituto Superiore di Sanità, effettua la sorveglianza delle caratteristiche cliniche dei casi nazionali attraverso apposito database, connesso alla piattaforma sopra citata.
I dati riferiti agli artt. 1 e 2 sono comunicati tempestivamente dall'Istituto Superiore di Sanità al Ministro della salute e, in forma aggregata, al Capo del Dipartimento della Protezione Civile e messi a disposizione delle Regioni e delle Province autonome di Trento e Bolzano.

Figura 3 riassume graficamente i flussi prima descritti.

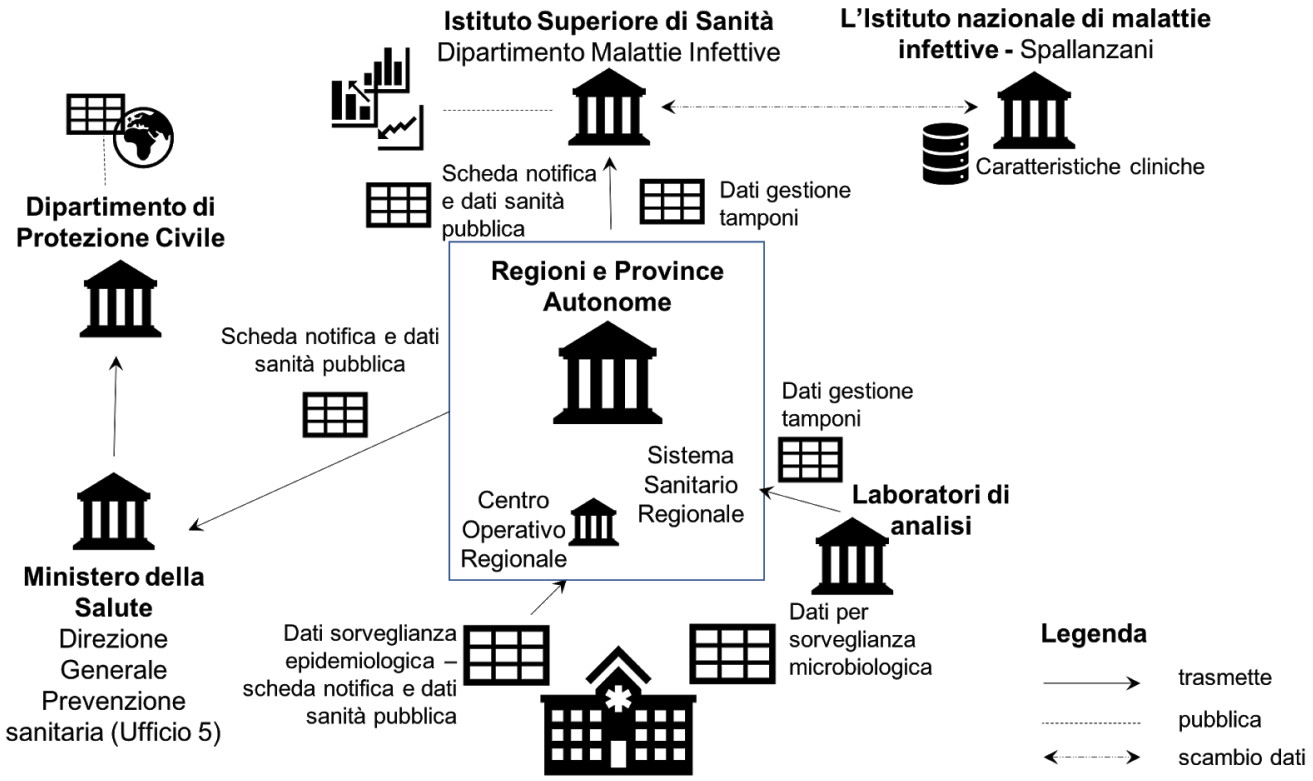


Figura 3: flussi informatici- notifica malattie infettive

4.5 Analisi dei consumi: dispositivi medici e medicinali

Partendo dai flussi individuati nella FASE A con priorità A, e coinvolti anche negli scenari di analisi della FASE B, si è avviato uno studio dei tracciati record di quei flussi di dati che ricadevano sotto il cappello di "analisi dei consumi/dei farmaci".
I tre flussi coinvolti sono

- DISPMED
- MEDOSP
- TRACCIA

I flussi sono stati analizzati in base alle sole informazioni a oggi disponibili, ossia i tracciati XSD e le specifiche funzionali individuate attraverso i riferimenti a pagine web del sito del Ministero della Salute.

4.5.1 MEDOSP

Per il flusso MEDOSP si trovano sia specifiche funzionali relative al tracciato record (file PDF), sia lo schema XSD del tracciato stesso. Inoltre, sono pubblicati manuali per guidare l'utente nel contesto del sistema NSIS.
In sostanza, il flusso MEDOSP è caratterizzato da dataset XML caricati dai vari referenti regionali all'interno del sistema NSIS previa autenticazione al sistema. Il caricamento consiste in un'operazione di upload di un file XML che viene validato dal sistema rispetto al tracciato fornito.
La semantica del tracciato è descritta in alcuni documenti PDF.
Il flusso contiene i dati sui consumi di diversi tipi di medicinali, con le relative quantità e costi, a opera di strutture sanitarie.

4.5.2 DISPMED

Per il flusso DISPMED si trovano informazioni sui tracciati record inseriti in alcuni decreti.
In particolare, vengono individuati:

- gli acquisti di dispositivi medici con riferimenti contrattuali e procedurali di approvvigionamento (e.g., dati sul CIG e il tipo di procedura adottata);
- i consumi dei dispositivi medici presso le strutture sanitarie

Non è stato trovato al momento lo specifico tracciato XSD.

4.5.3 TRACCIA

Per il flusso TRACCIA si trovano sia specifiche funzionali (file PDF), sia lo schema XSD dei tracciati. Il contenuto informativo riguarda la trasmissione dei dati relativi alle movimentazioni di farmaci in uscita da:

- produzione verso distribuzione all'ingrosso;
- produzione verso strutture del Sistema Sanitario Nazionale;
- produzione verso esercizi commerciali;
- distribuzione all'ingrosso verso altra distribuzione all'ingrosso;
- distribuzione all'ingrosso verso farmacie;
- distribuzione all'ingrosso verso SSN;
- distribuzione verso esercizi commerciali.

Il Gruppo di lavoro 2 ha predisposto un'analisi più approfondita del flusso, mirata ad analizzare il contesto infrastrutturale utilizzato per poter raccogliere i dati del flusso TRACCIA, i tracciati record utilizzati e i riferimenti open data dei dataset relativi ai flussi dei medicinali.

4.5.4 MEDOSP-DISPMED

In un tentativo di mettere insieme le strutture di due, al momento, dei tre flussi prima descritti e quindi individuare possibili elementi di collegamento tra i dati, si è predisposto un modello unitario ricavato da un'operazione di reverse engineering partendo dalle specifiche scritte in linguaggio naturale (file PDF) e i tracciati XSD. Il diagramma è riportato in Figura 4.

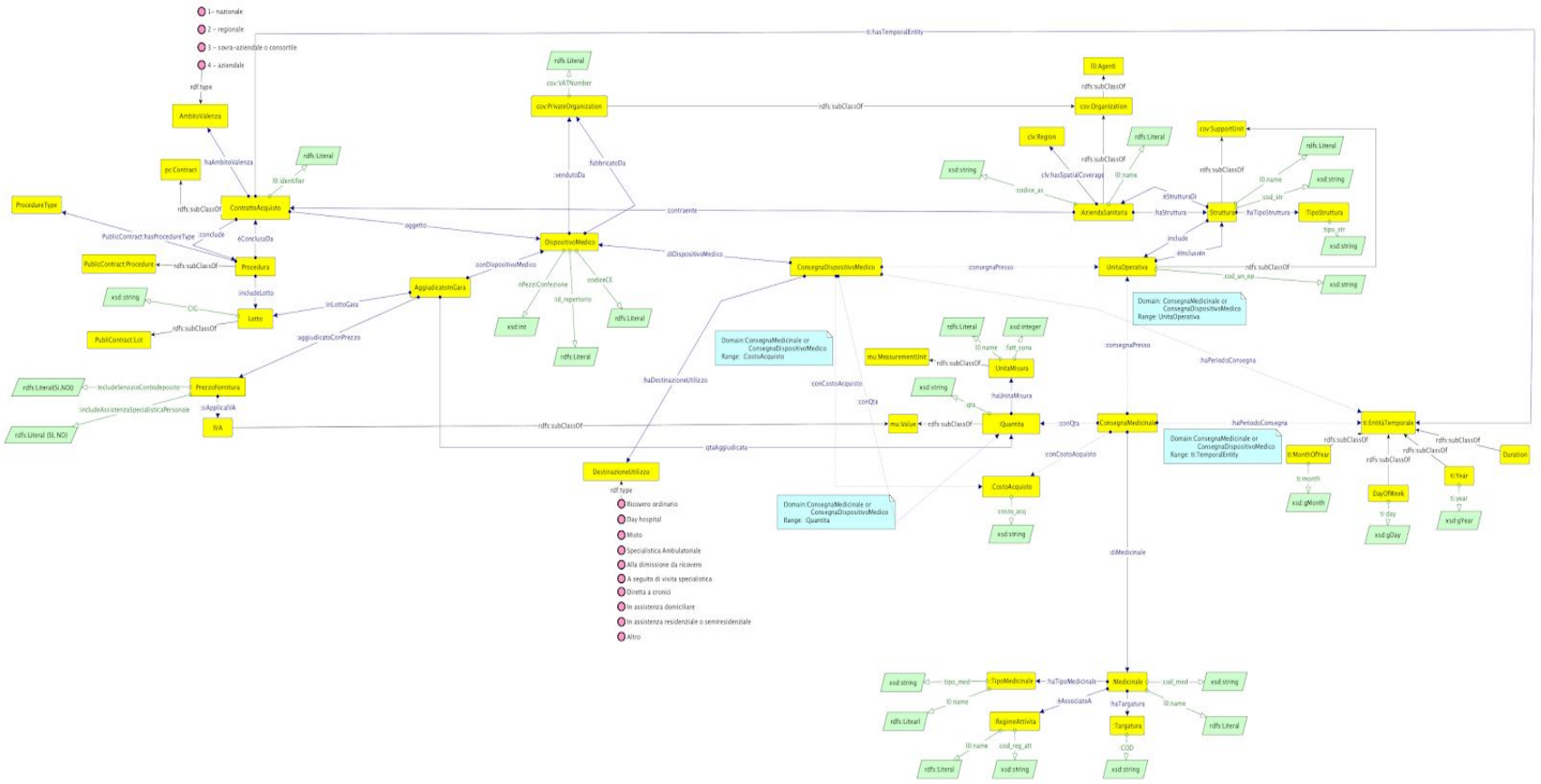


Figura 4: Modello dati per DISPMED e MEDOSP

I dataset, infatti, sono spesso trattati separatamente, anche rispetto a contesti diversi da quello in esame (esempio: contesto geografico, anagrafica di riferimento di istituzioni pubbliche, ecc.). Queste rende più complesso il processo di integrazione dei dati, necessario per poter inferire nuova conoscenza ed effettuare analisi più evolute su un più vasto patrimonio informativo Il vantaggio del lavoro presentato nel modello di Figura 4 è proprio quello di agevolare l'integrazione dei dati creando un livello unitario semantico (livello di interoperabilità semantica secondo i livelli

dell'European Interoperability Framework) che possa essere efficacemente sfruttato per collegare più facilmente i dati, grazie a standard che includono già nativamente tale possibilità.
Il modello è stato predisposto usando un tool grafico di disegno di ontologie (si definisce ontologia una specifica formale ed esplicita di rappresentazione/concettualizzazione condivisa di un dominio di conoscenza, definita sulla base di requisiti specifici). Il tool si chiama Graffoo è sviluppato dall'Università di Bologna e consente già, con appositi strumenti, di derivare il modello semantico dei dati espresso secondo gli standard del Web semantico, ampiamente usati nel mondo industriale e adottati anche a livello nazionale nello sviluppo della rete di ontologie e vocabolari controllati per la pubblica amministrazione.
I requisiti alla base del modello sono stati ricavati dai documenti di specifica trovati e sono state esplicitate le relazioni tra le diverse entità del dominio che nei documenti non erano specificatamente menzionate. Quest'ultimo punto implica che potrebbero essere state fatte delle assunzioni sulle relazioni non corrispondenti alla realtà di quello che viene tracciato nei domini di pertinenza dei due dataset.
Nel diagramma sono stati volutamente raggruppati elementi in giallo (classi) e in verde (attributi/campi) secondo questa logica:

- elementi di modellazione di contratti pubblici e di gare
- elementi sui dispositivi medici e la loro aggiudicazione nel contesto di contratti pubblici
- elementi sull'azienda sanitaria e le sue strutture
- elementi sulla consegna dei dispositivi medici (quantità e costi)
- elementi sui medicinali
- elementi sulla consegna dei medicinali (quantità e costi)
- elementi trasversali indipendenti dal dominio quali per esempio elementi temporali e geografici

Si noti che dove è stato possibile, gli elementi del modello sono stati allineati sia *direttamente* a ontologie nazionali (nel caso del tempo e dei luoghi), sia *indirettamente* (nel caso dei contratti pubblici, valori e unità di misura e organizzazioni).
Dal modello presentato in figura si nota come diversi elementi appartenenti a dataset diversi, anche trattati da uffici diversi dello stesso Ministero, possono essere messi in relazione tra loro al fine di abilitare sia analisi più complesse ma anche, grazie all'uso di standard semantici, inferenze di nuova conoscenza.

5. Spazio unico dati e risorse COVID-19

Nel contesto della task force, come progetto trasversale ai diversi gruppi di lavoro, sono stati raccolti dati e risorse (principalmente dashboard e atti normativi) pubblicati da attori istituzionali pubblici sull'emergenza COVID-19. L'obiettivo del lavoro, che si inquadra comunque in un'attività di raccolta di informazioni su dati ed elaborazioni degli stessi, è quello di fornire presso la task force uno spazio unico dove chiunque possa trovare bollettini, rapporti tecnici, mappe e dataset ufficialmente rilasciati da istituzioni pubbliche centrali e locali.
I contenuti della pagina https://dati-covid.italia.it/ sono stati classificati mediante l'uso di metadati obbligatori del profilo nazionale di metadatazione DCAT-AP_IT, usati per costruire la parte di presentazione della pagina. Non tutte le istituzioni sono rappresentate. A tal proposito si è scelto di pubblicare risorse che non fossero un mero elenco puntato di numeri in pagine web di notizie di siti istituzionali, ma piuttosto veri e propri dataset o elaborazioni più o meno elaborate di dati.
Nella mappatura si è data anche enfasi alla disponibilità secondo il paradigma open data, analizzando sia i formati di pubblicazione, sia le licenze d'uso adottate per la pubblicazione. Il più delle volte ci si è scontrati con formati chiusi (e.g., PDF) e licenze d'uso generali dei siti web istituzionali dove tutti i contenuti del sito, inclusi dati e infografiche risultavano coperti da copyright, tranne per pochi casi virtuosi.

6. Prime raccomandazioni

Dalle discussioni emerse dal Gruppo di lavoro 2 si possono proporre le prime seguenti raccomandazioni:

1. l'uso di codici identificativi per alcuni elementi di collegamento tra i dati (e.g., dati territoriali) è da incentivare maggiormente nella formazione delle basi di dati e si raccomanda che anche i flussi emergenziali utilizzino le codifiche nazionali delle anagrafiche nazionali ISTAT/NSIS;
2. la definizione di diversi flussi paralleli a quelli standard per la sola emergenza potrebbe non essere efficace in quanto può causare un eccessivo carico sui territori già impegnati nel salvataggio di vite umane. Il miglioramento dei

flussi standard con l'introduzione di una gestione opportuna per gli elementi di carattere emergenziale potrebbe essere preferibile. Questo implica la necessità di incentivare l'adozione del modello di interoperabilità tra i diversi sistemi informativi locali e centrali;

3. è bene specificare metadati di qualità per i dataset che vengono rilasciati in fase di emergenza giornalmente. A tal proposito ci sono già modelli standard del W3C che possono essere adottati;
4. il dato nella fase di emergenza è spesso incompleto, soggetto a cambiamenti che possono arrivare di giorno in giorno dai vari territori. A tal riguardo si raccomanda di affiancare alla pubblicazione giornaliera del dato la pubblicazione della serie storica stabilizzata, anch'essa metadatata con standard come quelli suggeriti nella raccomandazione precedente.
5. adottare standard comuni di modellazione allineandoli a quelli nazionali è raccomandabile per poter abilitare maggiori connessioni tra i dati e inferire nuova conoscenza;
6. si rafforza la raccomandazione delle linee guida nazionali per la valorizzazione del patrimonio informativo pubblico di mantenere il colloquio tra il livello centrale e locale, mediante interscambio automatico di dati aperti aggiornati, interrogabili attraverso API aperte che seguono il modello di interoperabilità. In tal senso, è raccomandato identificare l'insieme minimo di dati da pubblicare a livello centrale, anche secondo quanto stabilito da disposizioni normative, e quelli che il livello locale può ulteriormente dettagliare per cogliere le specificità della propria realtà locale;
7. per garantire sostenibilità nel tempo nell'apertura dei dati, si raccomanda di non gestire i dati aperti, fondamentali in tanti domini (e.g., salute, trasporti, lavoro, turismo, ecc.) nella gestione dell'emergenza, attraverso processi separati rispetto al processo di gestione tradizionale del dato;
8. si raccomanda nella fase di emergenza di creare un'unità di "gestione dati" con diversi elementi (umani, strumentali e formativi) fortemente interdisciplinare, con un nucleo di coordinamento di livello strategico e dei sotto-nuclei più operativi/specialistici. L'unità dovrebbe essere pronta e organizzata anche in precedenza, dovrebbe indicare la strategia ad alto livello per poi implementare la strategia con monitoraggi continui a livello più operativo. L'unità dovrebbe pertanto lavorare fin dall'inizio dell'emergenza affinché i flussi di gestione dei dati tradizionali possano essere sfruttati efficacemente anche per supportare dati in emergenza caratterizzati da frequenze di aggiornamento *real time*/giornaliere;
9. per supportare la precedente raccomandazione, si suggerisce di considerare l'istituzione, nel contesto interno delle amministrazioni, di figure specifiche con forti competenze nella gestione complessiva del ciclo di vita dei dati in

parte già suggerite nelle linee guida nazionali per la valorizzazione del patrimonio informativo pubblico, veri e propri Chief Data Steward a supporto dell'attuale figura del Responsabile Transizione Digitale (RTD). Tali figure avrebbero il compito di coordinare e collaborare con le controparti per i) sbloccare il valore di interesse pubblico dei dati, ii) individuare i problemi più insidiosi o domande sfidanti da formulare avendo dati a disposizione che aiutino anche a prioritizzare gli interventi, iii) proteggere le informazioni potenzialmente sensibili e iv) suggerire a livelli strategici come agire sulla base delle risultanze di elaborazioni fatte sui dati.

10. nel caso di messa a disposizione di forme di visualizzazione dei dati (e.g., dashboard, mappe, ecc) da parte delle amministrazioni, si raccomanda di accompagnarle sempre con i dati grezzi da rilasciare secondo il paradigma open data (formato aperto e licenza aperta), nel rispetto delle normative vigenti in tema di protezione dei dati personali. Se il dato grezzo è di tipo dinamico, ossia con una frequenza di aggiornamento molto elevata, è bene rendere disponibile il dato aperto mediante l'ausilio di Application Programming Interface (API), prevedendo comunque una sua storicizzazione da scaricare anche in blocco (*download in bulk*). Si raccomanda, infine, di metadatare il dato aperto rilasciato secondo gli standard nazionali ed europei di riferimento per la sua documentazione nei relativi cataloghi dei dati nazionali ed europei.

7. Gruppo di lavoro

La stesura di questo documento è stata curata da:

Maria Claudia Bodino (MID/PCM - Dipartimento per la Trasformazione Digitale)
Giorgia Lodi (Agenzia per l'Italia Digitale)
Roberto Polli (MID/PCM - Dipartimento per la Trasformazione Digitale)
Giovanni Baglio (INMP)
Serena Battilomo (Ministero della Salute)
Armando Cirillo (Ministero della Salute)
Stefania Garassino (Ministero della Salute)
Fabio Pammolli (Politecnico di Milano)
Pierluigi Cara (Protezione Civile)
Umberto Rosini (Protezione Civile)