DOI: 10.26314/GARR-Conf21-proceedings-17

Integrazione dei Principi FAIR nel processo di ricerca biomedica: la creazione del Registro della Regione Sardegna

Alessandro Sulis[1], Vittorio Meloni[1], Cecilia Mascia[1], Franco Cappai[2], Caterina G. Carboni[2], Ernesto D'Aloja[3], Giorgio Fotia[1], Giuseppe Serra[2], Giovanni Sotgiu[4], Paolo Uva[1], Francesca Frexia[1], Gianluigi Zanetti[1]

[1]CRS4: Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna, [2]Sardegna Ricerche, [3]Università degli Studi di Cagliari, [4]Università degli Studi di Sassari

Abstract. Il Registro per la Ricerca Biomedica della Regione Sardegna favorisce il riuso e la condivisione dei risultati della ricerca, sia da parte di esseri umani, sia di sistemi automatici (machine-actionability), in linea coi Principi FAIR. Questo lavoro descrive l'architettura del Registro, le scelte progettuali e implementative e gli sviluppi futuri in termini di completamento e messa in funzione del sistema. Il Registro è uno dei pilastri portanti del Programma I FAIR, iniziativa volta a promuovere le buone pratiche nella raccolta e gestione dei dati biomedici nell'ambito degli studi clinici indipendenti

Keywords. Ricerca biomedica, FAIR, Registro, Interoperabilità Semantica, FAIR Data Point, Molgenis

Introduzione

La difficoltà a impiegare in altri contesti i dataset creati durante lo svolgimento di un trial clinico (Vines et al. 2014), così come la bassa riproducibilità dei risultati ottenuti (Freedman et al. 2015), hanno generato un interesse via via crescente verso l'Open Science (Allen et al. 2019) e verso la formalizzazione di linee guida come i Principi FAIR (Wilkinson et al. 2016), che promuovono la creazione nel processo di ricerca di digital research objects (dati, algoritmi, strumenti, etc.) che siano Findable, Accessible, Interoperable e Re-Usable.

Il Registro per la Ricerca Biomedica della Regione Sardegna è uno strumento pensato per affrontare queste criticità supportando la condivisione secondo i Principi FAIR di informazioni descrittive (metadati) sui dataset creati nell'esecuzione di studi clinici indipendenti, compatibilmente con i principi etici, le disposizioni legali e la protezione della proprietà intellettuale. La realizzazione del Registro è uno dei pilastri fondamentali del Programma I FAIR (Cappai et al. 2019), illustrato in Figura 1, l'iniziativa di formazione e trasferimento tecnologico che ha come obiettivo quello di diffondere e promuovere buone pratiche nella raccolta, gestione e condivisione di dati biomedici, da un punto di vista tecnologico, etico e statistico.

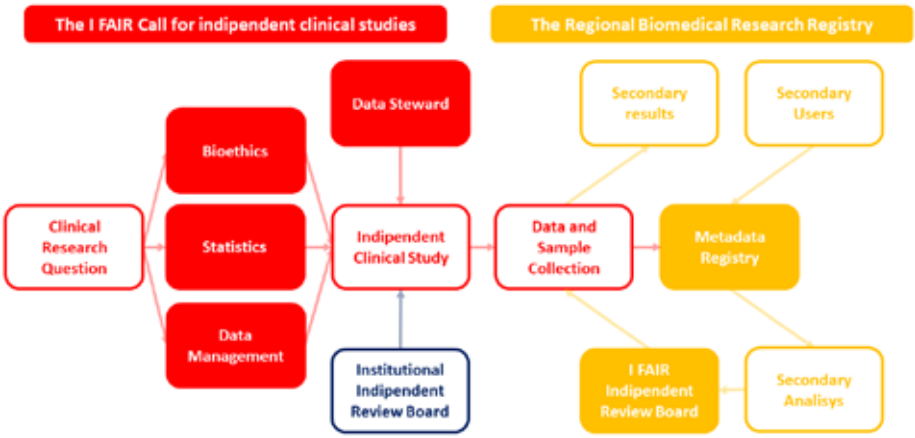


Fig. 1
Il programma IFAIR della Regione Sardegna

1. Il Registro

Il Registro è stato progettato e realizzato per essere una risorsa FAIR, ovvero una collezione di "oggetti" caratterizzata da contenuti e funzionalità acessibili e utilizzabili direttamente da esseri umani e sistemi automatici (machine-actionability). L'implementazione segue le principali best-practices per il processo di FAIRificazione (Jacobsen et al. 2020), attuate secondo le fasi rappresentate nella Figura 2.

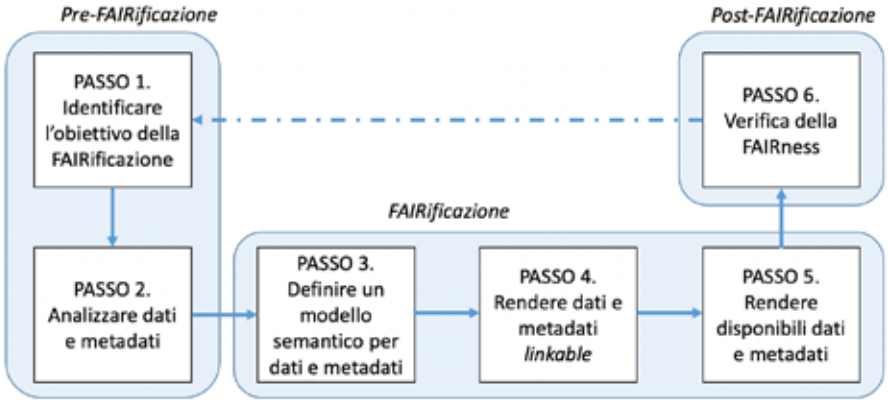


Fig. 2
Workflow di FAIRificazione, (riadattato da Jacobsen, 2020)

Il primo passo del workflow è la definizione dell'obiettivo del processo, che coincide con l'obiettivo operativo del Registro: la condivisione di informazioni aggregate e parametri descrittivi sugli studi clinici (metadati). Una prima selezione di metadati comuni ai vari studi è quindi stata analizzata (Passo 2) per definire la strategia per la fase di FAIRificazione vera e propria, che prevede: la definizione del modello dei concetti (modello semantico) descritti da dati e metadati (Passo 3), la loro trasformazione in un formato machine-readable (Passo 4), la realizzazione di uno strato di persistenza e condivisione (Passo 5). Il Registro implementa questi passi seguendo le specifiche FAIR Data Point

(FDP) (Da Silva Santos et al. 2016), che mettono a disposizione un Semantic Data Model, basato sull'ontologia Data Catalog Vocabulary (DCAT) (Maali et al. 2014), insieme a un set di API per popolare e interrogare un registro di metadati nel rispetto dei Principi FAIR. I metadati selezionati per gli studi sono stati quindi mappati, ove possibile, sul modello semantico fornito dalle specifiche del FDP. Per i metadati più specifici, come ad esempio "area terapeutica" e "tipologia di dati raccolti", per i quali non è stata trovata una corrispondenza diretta con i concetti del modello FDP, il modello stesso è stato esteso utilizzando la Semanticscience Integrated Ontology (SIO) (Dumontier et al. 2014). I relativi metadati sono stati quindi valorizzati mediante le principali ontologie in uso nel dominio clinico (e.g., ICD e MESH), in modo da definire ogni valore senza ambiguità, a supporto dell'interoperabilità semantica.

L'architettura e l'implementazione del Registro sono basate su Molgenis (van der Velde et al. 2019), un'applicazione open source per la raccolta, la gestione e la condivisione di dataset di ricerca. Molgenis implementa nativamente le specifiche del FDP e permette la creazione di interfacce grafiche per la consultazione e la gestione dei dataset. Nel contesto del Registro è stato esteso il modello semantico del FDP di Molgenis ed è stata creata un'interfaccia grafica per la ricerca e l'accesso ai metadati degli studi. Per quanto riguarda la fase di Post-FAIRificazione (Passo 6), sono stati predisposti alcuni test per valutare l'effettiva copertura del sistema rispetto ai Principi FAIR, sulla base di specifiche metriche (Wilkinson et al. 2019).

1.2 Stato del prototipo

La versione attuale del Registro prevede un'istanza di Molgenis apposita, all'interno della quale il modello FDP esteso è stato popolato con un primo insieme di metadati generici relativi ai 18 studi facenti parte del Programma I FAIR. Il prototipo è dotato di un'interfaccia utente per la ricerca degli studi sulla base di metadati di diverso tipo e per la consultazione delle informazioni di dettaglio relative a ciascuno studio. Per ogni parametro, la ricerca è ristretta ai valori possibili dello specifico concetto ontologico associato: ad esempio, per il parametro "area terapeutica" sarà possibile effettuare la ricerca solo sulla base dei valori relativi dell'ontologia MESH.

2. Conclusioni

Il Registro per la Ricerca Biomedica della Regione Sardegna è uno strumento realizzato secondo i Principi FAIR e finalizzato a raccogliere un insieme di metadati descrittivi relativi a studi clinici indipendenti, per facilitare la condivisione e il riuso dei dati raccolti. Il Registro è in fase di popolamento con i metadati degli studi selezionati dal Programma I FAIR, mentre in seguito verranno compresi anche i riferimenti ad altri studi che vorranno aderire.

Sviluppi futuri prevedono l'estensione del modello semantico del Registro, ad esempio per includere concetti che possano descrivere, quando disponibili, la provenance dei dati generati nel processo di ricerca, al fine di supportare un miglioramento della riproducibilità dei risultati. Il Registro può essere facilmente riadattato per raccogliere informazioni

descrittive riguardo a tipi diversi di studi di ricerca. Per favorire un riuso in altri contesti che richiedano la condivisione di informazioni su studi clinici secondo i Principi FAIR, il modello semantico è stato pubblicato (ifair-reg, 2021) e al termine dello sviluppo il codice verrà rilasciato in open source. L'intero approccio proposto dal Programma I FAIR, che combina le soluzioni tecnologiche con formazione e supporto multidisciplinare in un percorso verso una gestione e condivisione più "FAIR" dei risultati della ricerca, si candida ad essere considerato un modello di riferimento, in quanto l'analisi effettuata durante la preparazione del piano di attività non ha evidenziato la presenza di iniziative simili nel contesto di studi clinici indipendenti.

Ringraziamenti

I contributi degli autori sono realizzati nell'ambito del Programma I FAIR (finanziamento da Sardegna Ricerche, nell'ambito del POR FESR 2014/2020, Progetto FAIR DATA), del Progetto DIFRA (finanziamento dalla Regione Autonoma della Sardegna) e del Progetto EJP-RD (finanziamento H2020, Grant n. 825575).

Riferimenti bibliografici

C. Allen, D. M. Mehler (2019), Open science challenges, benefits and tips in early career and beyond, PLoS biology 17 (5).

F. Cappai, C. G. Carboni, E. D'Aloja, G. Fotia, F. Frexia, G. Serra, G. Sotgiu, P. Uva, G. Zanetti (2019), I FAIR Program: the Sardinian way to support and fund independent clinical studies that want to be Findable Accessible Interoperable Reusable, in: The Ecosystem of Evidence Conference - Abstract book, GIMBE, Evidence for Health, p. 15.

L. B. Da Silva Santos, M. D. Wilkinson, et al. (2016), FAIR data points supporting big data interoperability, Enterprise Interoperability in the Digitized and Networked Factory of the Future. ISTE, London, pp. 270–279.

M. Dumontier, C. J. Baker, et al. (2014), The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery, Journal of biomedical semantics 5 (1), pp. 1–11.

L. P. Freedman, I. M. Cockburn, T. S. Simcoe (2015), The economics of reproducibility in preclinical research, PLoS Biol 13 (6) e1002165.

A. Jacobsen, R. Kaliyaperumal, et al. (2020), A generic workflow for the data FAIRification process, Data Intelligence 2 (1-2), pp. 56–65.

ifair-reg (2021). URL: https://github.com/crs4/ifair-reg.

F. Maali, J. Erickson, P. Archer (2014), Data catalog vocabulary (DCAT), W3C recommendation 16.

K. J. van der Velde, F. Imhann, et al. (2019), MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians, Bioinformatics 35 (6), pp. 1076–1078.

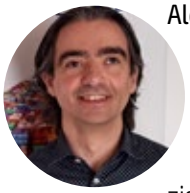
T. H. Vines, A. Y. Albert, et al. (2014), The availability of research data declines rapidly with article age, Current biology 24 (1), pp. 94–97.

M. D. Wilkinson, M. Dumontier, et al. (2016), The FAIR guiding principles for scientific

data management and stewardship, Scientific data 3 (1), pp. 1–9.

M. D. Wilkinson, M. Dumontier, et al. (2019), Evaluating FAIR maturity through a scalable, automated, community-governed framework, Scientific data 6 (1), pp. 1–12.

Autori

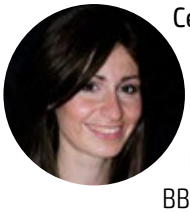


Alessandro Sulis alessandro.sulis@crs4.it

Alessandro Sulis è un ricercatore del gruppo Digital Health del CRS4 dal 2007. Ha un'esperienza decennale nell'interoperabilità tra sistemi clinici secondo i principali standard e best practices e collabora attivamente con gli organismi internazionali di riferimento (IHE e HL7 International). I suoi attuali interessi di ricerca comprendono lo studio e la promozione dei Principi FAIR, nel contesto di progetti legati alla FAIRificazione di dati e metadati in ambito clinico.

Vittorio Meloni vittorio.meloni@crs4.it

Dalla laurea in Informatica ottenuta presso l'Università degli Studi di Cagliari nel 2008, si occupa di ricerca e sviluppo nel campo dell'Informatica Clinica e delle Applicazioni Biomediche. Lavora dal 2010 nel gruppo Digital Health del CRS4, nel quale è impegnato in progettazione e sviluppo di software focalizzati su integrazione di domini clinici tramite standard e linee guida internazionali, condivisione di dati della ricerca secondo i Principi FAIR e telemedicina in tempo reale.



Cecilia Mascia cecilia.mascia@crs4.it

Cecilia Mascia, ingegnere biomedico, dal 2014 fa parte del gruppo Digital Health del CRS4, occupandosi principalmente dello sviluppo di modelli computabili di dati clinici mediante standard e ontologie internazionali (e.g., openEHR, HL7, OMOP), orientati all'interoperabilità semantica e al riuso dei dati in ottica FAIR. Attualmente collabora con il consorzio BBMRI-ERIC alla stesura di un proposal per lo Standard ISO 23494 dedicato alla cattura della provenance di dati e campioni biologici.

Franco Cappai cappai@sardegnaricerche.it

Franco Cappai fa parte dell'Unità di Supporto alla Ricerca Biomedica di Sardegna Ricerche dove coordina il Programma I FAIR. Ha un'esperienza ventennale nel trasferimento tecnologico e nella valorizzazione della ricerca con specializzazione in ambito biomedico. Attualmente partecipa al gruppo di lavoro ELSI di BBMRI-IT sul "Biobancaggio traslazionale" ed è patient expert reviewer per BMJ e NIHR.

Caterina G. Carboni carboni@sardegnaricerche.it

Caterina G. Carboni, laurea in Chimica e Tecnologie Farmaceutiche, dopo essersi occupata della gestione tecnico scientifica dei progetti in sviluppo preclinico per una società di sviluppo di nuovi agenti diagnostici e terapeutici, dal 2016 collabora con Sardegna Ricerche nell'organizzazione della Filiera Biomed occupandosi prevalentemente dell'introduzione dei Principi FAIR nella ricerca clinica spontanea ed indipendente e dell'istituzione del Registro della Ricerca Biomedica in Sardegna.

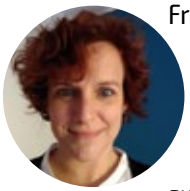
Ernesto D'Aloja ernestodaloja@gmail.com
Ernesto D'Aloja è Professore Ordinario di Medicina Legale nella Facoltà di Medicina e Chirurgia dell'Università di Cagliari e Direttore della Scuola di Specializzazione in Medicina Legale della stessa Università. È autore di oltre 140 pubblicazioni (I.F.>114) su riviste nazionali ed internazionali in tema di genetica forense, responsabilità professionale medica, patologia e tossicologia forense, e bioetica.

Giorgio Fotia giorgio.fotia@crs4.it
Giorgio Fotia è il direttore del Programma di Bioinformatica del CRS4. Il suoi attuali interessi di ricerca si concentrano sui metodi numerici ad alte prestazioni per la simulazione di problemi di biologia computazionale, sull'integrazione e sullo sviluppo di tecnologie abilitanti per l'analisi di dati biomedicali, e sulla loro applicazione alla soluzione di problemi data-intensive delle bioscienze. Attualmente è vicepresidente della Società Italiana di Matematica Industriale e Applicata ed Editor in Chief di Communications in Applied and Industrial Mathematics.

Giuseppe Serra giuseppe.serra@sardegnaricerche.it
Giuseppe Serra è responsabile del settore Startup e sviluppo del Parco scientifico e tecnologico dell'agenzia regionale Sardegna Ricerche. Al settore fanno capo le attività di supporto alla creazione e sviluppo di startup innovative e spin off dalla ricerca, la partecipazione a diversi programmi UE in materia di innovazione e trasferimento tecnologico e la gestione di programmi della Piattaforma biomed del parco e dell'Unità di supporto alla ricerca biomedica. Nel corso della sua esperienza ha ricoperto il ruolo di amministratore presso alcuni centri di ricerca regionali.

Giovanni Sotgiu gsotgiu@uniss.it
Giovanni Sotgiu è professore ordinario presso l'Università degli Studi di Sassari nelle tematiche di statistica medica, metodologia statistica applicata in ambito biomedico e clinico, organizzazione sanitaria, management e gestione sanitaria, informatica. Ha svolto e svolge ruolo in progetti di ricerca nazionali ed internazionali. Membro di diverse società scientifiche nazionali ed internazionali e di board editoriali di riviste internazionali.

Paolo Uva paolouva@gaslini.org
Paolo Uva, precedentemente responsabile della Piattaforma di Next Generation Sequencing del CRS4 è responsabile del gruppo di Bioinformatica Clinica presso l'IRCCS Gaslini (Genova) da maggio 2020. Le sue attività di ricerca sono centrate sull'analisi dei dati generati con la tecnologia del sequenziamento massivo e l'integrazione con altre tecnologie omiche. Le principali applicazioni riguardano lo studio delle basi genetiche delle malattie ereditarie.



Francesca Frexia francesca.frexia@crs4.it
Francesca guida il Programma di Ricerca Digital Health del CRS4. I suoi presenti interessi di ricerca si concentrano su modellazione di dati e processi biomedici e applicazione dei Principi FAIR, con un'attenzione particolare agli aspetti di interoperabilità e tracciabilità. Sta attualmente lavorando sulla modellazione dati con il formalismo openEHR e sulla stesura di linee guida e standard per la tracciabilità del campione (IHE SET Profile) e la provenance nelle biotecnologie (ISO 23494).

Gianluigi Zanetti gianluigi.zanetti@crs4.it

Gianluigi Zanetti è stato il Direttore del settore Data-intensive Computing del CRS4. Le sue attività di ricerca hanno toccato molte aree della scienza computazionale, tra cui la meccanica statistica computazionale e la simulazione del flusso sanguigno. Negli ultimi anni, si è dedicato in particolare allo sviluppo di strategie e infrastrutture in grado di garantire, in modo completamente tracciabile e riproducibile, l'elaborazione e l'analisi di grandi quantità di dati eterogenei e complessi.