HemaFAIR: Implementing FAIR principles in real-world Rare Hematological Disease datasets

Stella Tamana[1], Christina Yiangou[1], Kalia Orphanou[1], Maria Xenophontos[1], Panayiota L. Papasavva[1], César Bernabé[2], Marco Roos[2], Julio Marchiori Dias[2], Martijn G. Kersloot[3], Ronald Cornet[3], Anna Minaidou[1], Coralea Stephanou[1], Annalisa Landi[4], Viviana Giannuzzi[4], Fedele Bonifazi[4], Carsten W. Lederer[1] and Petros Kountouris[1,*]

[1]*Department of Blood Disorder Genetics & Thalassemia, The Cyprus Institute of Neurology & Genetics, 6 Iroon Avenue, 2371 Agios Dometios, Nicosia, Cyprus*

[2]*Department of Human Genetics, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands*

[3]*Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Amsterdam, 1105 AZ, The Netherlands*

[4]*Fondazione per la ricerca farmacologica Gianni Benzi onlus, Bari, Italy*

Abstract

Data fragmentation and lack of standardization in Rare Hematological Diseases (RHDs) hinder research and delay patient care by limiting the reusability of clinical information. The HemaFAIR project addresses these challenges by applying the FAIR principles to establish a robust, patient-centered data ecosystem. This work reports on the initial implementation of the FAIR principles in two use cases: the Cyprus Haemoglobinopathy Patient Registry and the INHERENT platform. We demonstrate the successful application of a stepwise FAIRification workflow and semantic modeling to enhance registry value and interoperability.

Keywords

Rare Disease, GO-Plan, FAIR, Patient Registry

1. Introduction

Rare Hematological Diseases (RHDs) are characterized by small, geographically dispersed patient populations. Research in this field is significantly impeded by data fragmentation and a lack of standardization across isolated registries. To address this, we apply the Findable, Accessible, Interoperable, and Reusable (FAIR) principles to two key use cases: the national Cyprus Haemoglobinopathy Patient Registry (CYHAPR) and the INHERENT platform [1]. Our objective is to transform these silos into a patient-centered data ecosystem capable of supporting secure, cross-registry research.

2. Methods

Our FAIRification strategy began with the Goal-Oriented Planning method [2], identifying six central competency questions (CQs) to align technical steps with clinical needs. Using a stepwise, CQ-driven

SWAT4HCLS: 17th International SWAT4HCLS Conference, March 23–26, 2026, Amsterdam, The Netherlands

*Corresponding author.

stellat@cing.ac.cy (S. Tamana); christinay@cing.ac.cy (C. Yiangou); kaliao@cing.ac.cy (K. Orphanou); mxenoph@cing.ac.cy (M. Xenophontos); panayiotap@cing.ac.cy (P. L. Papasavva); c.h.bernabe@lumc.nl (C. Bernabé); m.roos@lumc.nl (M. Roos); j.c.marchiori_dias@lumc.nl (J. Marchiori Dias); m.g.kersloot@amsterdamumc.nl (M. G. Kersloot); r.cornet@amsterdamumc.nl (R. Cornet); annami@cing.ac.cy (A. Minaidou); coraleas@cing.ac.cy (C. Stephanou); al@benzifoundation.org (A. Landi); vg@benzifoundation.org (V. Giannuzzi); fb@benzifoundation.org (F. Bonifazi); lederer@cing.ac.cy (C. W. Lederer); petrosk@cing.ac.cy (P. Kountouris)

0000-0002-0877-7063 (S. Tamana); 0009-0002-2230-0461 (C. Yiangou); 0000-0003-1351-2489 (K. Orphanou); 0000-0001-5978-0193 (M. Xenophontos); 0000-0003-4410-3119 (P. L. Papasavva); 0000-0003-1795-5930 (C. Bernabé); 0000-0002-8691-772X (M. Roos); 0009-0003-2347-6490 (J. Marchiori Dias); 0000-0002-8691-772X (M. G. Kersloot); 0000-0002-1704-5980 (R. Cornet); 0000-0001-9078-7772 (A. Minaidou); 0000-0002-2546-6478 (C. Stephanou); 0000-0001-9368-6424 (A. Landi); 0000-0002-2534-6783 (V. Giannuzzi); 0000-0003-4935-5702 (F. Bonifazi); 0000-0003-3920-9584 (C. W. Lederer); 0000-0003-2681-4355 (P. Kountouris)

workflow in Protégé [3], we modeled the initial CQ, specifically modeling thalassemia genotypes (stratified by centre, region, country, and ethnic group) and their association with transfusion status. We utilized Ontop within Protégé to materialize relational data into RDF, validating the resulting graph in GraphDB via SPARQL querying and visual inspection.

Simultaneously, to evaluate the coverage of existing models against our specific research needs, we initiated mappings to both the OMOP Common Data Model (CDM) and the CARE Semantic Model (CARE-SM). This parallel workflow allows us to assess whether a standardized schema is sufficient or if a custom model is necessary to capture the granularity of our data. As the mapping to both OMOP CDM and CARE-SM progresses, we maintain a rigorous map of concepts and ontologies specific to each model for all modeled parameters.

3. Results & Discussion

The initial semantic model successfully addressed the primary CQ, establishing a foundation for the remaining questions. To enhance findability, a FAIR Data Point (https://w3id.org/HemaFAIR/fdp) [4] was deployed for exposure of dataset-level metadata. Both CYHAPR and INHERENT were integrated into the European Rare Disease Registry Infrastructure (ERDRI). This integration encompassed registration in ERDRI.dor and the deposition of metadata in ERDRI.mdr. Furthermore, the ERDRI.spider pseudonymization service was implemented for CYHAPR to enable GDPR-compliant data linkage.

We demonstrated the value of a stepwise, CQ-driven approach for the retrospective FAIRification of RHD platforms. Future work will focus on the FAIRification of Patient-Reported Outcome Measures.

Acknowledgments

This project has received funding from the European Union's Horizon Europe Twinning Bottom-Up Programme under Grant Agreement No. 101159589.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini for grammar and spell check. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Kountouris, C. Stephanou, N. Archer, F. Bonifazi, V. Giannuzzi, K. H. M. Kuo, A. Maggio, J. Makani, M. D. M. Mañú-Pereira, K. Michailidou, S. Nkya, O. E. Nnodu, S. Trompeter, L. Tshilolo, A. Wonkam, B. A. Zilfalil, B. P. D. Inusa, M. Kleanthous, on behalf of the International Hemoglobinopathy Research Network (INHERENT), The international hemoglobinopathy research network (INHERENT): An international initiative to study the role of genetic modifiers in hemoglobinopathies, Am. J. Hematol. 96 (2021) E416–E420. doi:10.1002/ajh.26323.

[2] C. Bernabé, I. Valle, T. Prince Sales, E. Schultes, N. van Ulzen, V. E. Silva Souza, A. Jacobsen, L. O. Bonino da Silva Santos, B. Mons, M. Roos, GO-plan: A goal-oriented method for FAIRification planning, Inf. Serv. Use 45 (2025) 100–124.

[3] M. A. Musen, Protégé Team, The protégé project: A look back and a look forward, AI Matters 1 (2015) 4–12. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4883684/. doi:10.1145/2757001.2757003.

[4] L. O. B. da Silva Santos, K. Burger, R. Kaliyaperumal, M. D. Wilkinson, FAIR data point: A FAIR-oriented approach for metadata publication, Data Intell. 5 (2023) 163–183.