

A modular approach to knowledge graphs and FAIR data in healthcare

Matthijs Sloep ^[0000-0003-3602-1885], Petros Kalendralis, Johan van Soest, Rianne Fijten ^[0000-0002-1964-6317]

Department of Radiation Oncology (MAASTRO), GROW school for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands
matthijs.sloep@maastro.nl

Keywords: FAIR data, Linked Data, Modular, knowledge graphs

1 Introduction

In healthcare, and more specifically cancer treatment, data sharing is essential yet difficult. 1 in 5 people diagnosed with cancer have a rare type of cancer, which means considerable time is needed to collect sufficient data for research. Combining data from multiple centres is therefore vital, unfortunately, linking this data is not straightforward. There are various ways healthcare centres store their data, due to for instance differences in treatment protocols and clinical systems. This means different variables and annotations are used. Consequently before we can solve any medical problems, we first need to solve this data integration challenge.

The FAIR (Findable, Accessible, Interoperable and Reusable) principles [1] facilitate data driven research and integration. Here we describe a semantic solution to the application of the FAIR principles in a clinical use case and address the challenges we faced and the practical solutions we implemented to solve these issues. Our use case is proTRAIT, a nationwide Proton beam therapy (PBT) data infrastructure which combines data from the various proton centres in the Netherlands into a centralized database.

Our approach to the FAIRification process is akin to the steps described by the Go-FAIR initiative: collect the data, analyse the data, define a semantic model for the data, link the data, define the meta-data, and finally deploy a resource for the data. We started with analysing what data is collected at the treatment centres and is relevant for medical research. Radiotherapists from the Dutch national proton therapy platform listed these specific clinical and treatment items. Then we manually annotated all the different items and created a knowledge graph to interconnect them. The list items were defined with existing and new ontology classes using the Radiation Oncology Ontology [2] The non-FAIR relational data from the various centres will be converted into triples utilizing the knowledge graphs. Essentially this process creates findable, accessible and linked semantic data. Metadata is collected and described in the dataset and the use of ontologies and unique and persistent identifiers ensure the interoperability. Finally, the reusability will be cared for by hosting the data on a clinical data management tool.

Our approach means that a lot of work goes into creating and maintaining the knowledge graphs. Each cancer type requires a separate, manually created knowledge graph, and each list and graph consist of hundreds of items. To facilitate this process we made full use of the considerable overlap between the respective lists: by creating separate turtle files for small subsets of items we leveraged the modular characteristics of knowledge graphs. We designed these files in such a way that they can be easily reused. Apart from the technical composition of the code another challenge was to arrange these items in subsets that are both logical and practical, in order for the selection of the necessary subsets to be as easy as possible. For each separate cancer we selected relevant turtle files, created more if necessary and compiled a tailored comprehensive graph. Another advantage of this modular approach is that we can easily adapt and adjust our graphs to changes in the clinic. For instance, when new items prove relevant to a research question we can easily add more variables.

References

1. Wilkinson, MD, Dumontier, M, Aalbersberg, IJJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, JW, da Silva Santos, LB, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJG, Groth, P, Goble, C, Grethe, JS, Heringa, J, 't Hoen, PAC, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, ME, Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, SA, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, MA, Thompson, M, Van Der Lei, J, Van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J & Mons (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data*, 15 (3) Article 160018, 10.1038/sdata.2016.18
2. Traverso A, van Soest J, Wee L, Dekker A. (2018) The Radiation Oncology Ontology (ROO): publishing linked data in radiation oncology using semantic web and ontology techniques. *Med Phys*. 10.1002/mp.12879.