The Knowledge Graph Lifecycle in NTT DATA ⋆

Javier Flores[1,*], Emmanuel Jamin[2], Sergi Nadal[1] and Oscar Romero[1]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain
[2]NTT Data, Barcelona, Spain

1. Introduction

The Semantic Business Unit (SEMBU) in NTT DATA aims to increase the semantic interoperability and accessibility of European institutions' data projects by following Linked Open Data (LOD) principles to build controlled vocabularies and produce Knowledge Graphs (KGs). One of its most notable projects revolves around the CORDIS portal[1], which publishes information about research and innovation projects funded by the European Commission. SEMBU pursues two main goals: *(i)* expose semantic data related to CORDIS via a SPARQL endpoint that facilitates access and reuse of quality scientific-related data, and *(ii)* design an efficient, incremental, and automated KG lifecycle to be used as a reference in other data projects. To that end, we have adopted state-of-the-art semantic technologies to support the creation and management of the KG with the goal of centralizing knowledge and providing an overall view of data assets that improve data governance, maintenance, and external interaction by data consumers. We have also identified some of their limitations which are tackled via an industrial PhD. This paper reports our experience, the obstacles, and proposals for generating and maintaining the CORDIS KG.

2. CORDIS KG lifecycle

CORDIS KG is generated using the lifecycle depicted in Figure 1, which considers the business needs and incremental integration needs with legacy systems. The remainder of the section is devoted to describing such lifecycle.

Collection. XML files are collected from the CORDIS portal and materialized into RDF triples. Each file is automatically analyzed and mapped to resources of the EURIO ontology[2] via RML mappings. Since files evolve in their schema and data, manually updating the mappings is a laborious and error-prone task. We thus proposed a largely automated process bootstrapping the file schema and automatically updating the RML mappings.

The 21st International Semantic Web Conference, October 23–27, 2022, Hangzhou, CN

⋆This work was partly funded by the Spanish Ministerio de Ciencia e Innovación under project PID2020-117191RB-I00 (DOGO4ML). Javier Flores is supported by contract 2020-DI-027 of the Industrial Doctorate Program of the Government of Catalonia and CONACYT's scholarship. Sergi Nadal is partly supported by the Spanish Ministerio de Ciencia e Innovación, as well as the European Union - NextGenerationEU, under project FJC2020-045809-I / AEI/10.13039/501100011033.

*Corresponding author.
jflores@essi.upc.edu (J. Flores); emmanueljeanjacques.jamin@nttdata.com (E. Jamin); snadal@essi.upc.edu (S. Nadal); oromero@essi.upc.edu (O. Romero)

[1]cordis.europa.eu
[2]op.europa.eu/en/web/eu-vocabularies/eurio

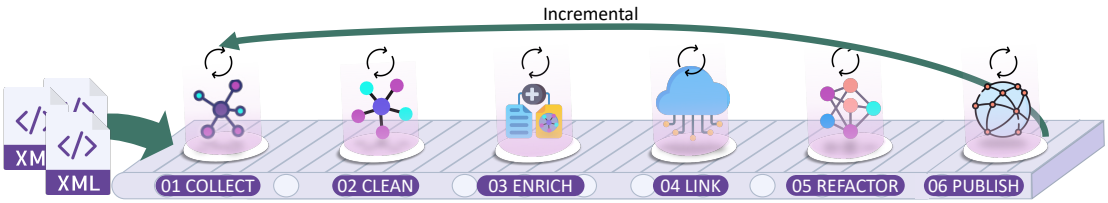

Figure 1: Overview of the implemented KG lifecycle in CORDIS.

Clean. A specific set of quality rules and SHACL shapes act as entity resolution system. Inconsistencies and errors are solved by manual validation, determining the best strategy to solve the conflict. Due to the manual effort, using learning techniques that combine attribute-level data could improve the detection of similarities and discrepancies between entities and suggest resolution strategies.

Enrich. Unstructured data related to the project funding is analyzed using named entity recognition tools to identify relevant information (e.g., organisations and people) to link them to EURIO resources. Here, CRF-NER[3] has facilitated this step due to its generic infrastructure, allowing to reuse already created resources within the KG and efficiently enriching them.

Link. LOD repositories are explored via schema and instance alignment tools (e.g., LogMap, Alignment API, and AML) to enhance the added value of the KG. However, the low precision of alignment tools due to lexical similarity bias (e.g., syntactically similar elements which are semantically different), and their inability to scale up to large volumes of data has lead to manual approaches. To overcome these limitations, graph data profiles can be compared using learning techniques to predict their expected similarity and reduce the low precision.

Refactor. CORDIS KG is manually updated using hints acquired in previous steps, leading to different KG versions. Tracing changes made in the schemata and instances is crucial for supporting evolution transparency between versions. Thus, we propose a versioning mechanism focused on entities and individual changes inspired by the PAV ontology[4]. Moreover, automated refactoring hints via learning techniques is an exciting direction that could maximize the linking of LOD repositories and reduce human effort.

Publish. Finally, the publication of the enriched KG consists of *(i)* ensuring best practices for publishing ontologies on the web using FOOPS!, and *(ii)* generating the proper documentation using WIDOCO. Adhering to the previously described lifecycle, the Publication Office of the European Union plans to release the first version of the CORDIS KG through a SPARQL endpoint in December 2022.

In the on-site presentation, we will present each of the phases, the tools used and the level of automatization that is possible to achieve. The complete lifecycle will be exemplified with real data samples.

[3] nlp.stanford.edu/software/CRF-NER.html
[4] pav-ontology.github.io/pav/