

# Ontology-based data integration and open data publication: the case of the Italian Arthroplasty Registry

Riccardo Valentini<sup>1,2</sup>, Eugenio Carrani<sup>1</sup>, Duilio Luca Bacocco<sup>1,2</sup>, Marina Torre<sup>1</sup>, Maurizio Lenzerini<sup>2</sup>

<sup>1</sup> Scientific Secretariat of the Presidency, Italian National Institute of Health, Rome, Italy  
<sup>2</sup> Department of Computer, Control and Management Engineering, Sapienza University of Rome

[valentini@diag.uniroma1.it](mailto:valentini@diag.uniroma1.it)

## INTRODUCTION

The Italian Arthroplasty Registry (RIAP) data collection system is organized on a federal basis, involving 13 Italian regions (most of them, with their own data collection methodology).

Aim of this study is to define a methodology to execute data integration tasks automatically, so that regional data might flow to a central semantic-based repository, according to a common standard, a first step towards an ontology-based open data publication.

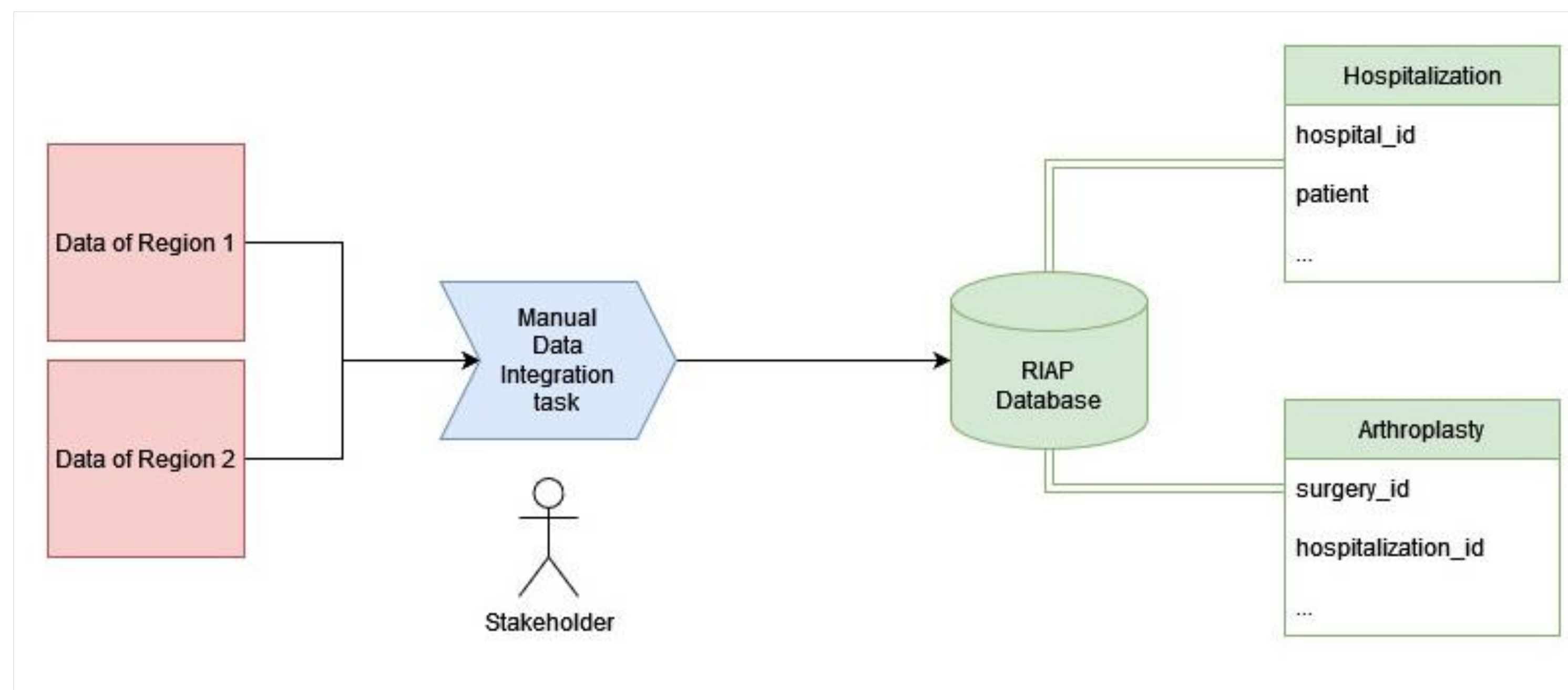
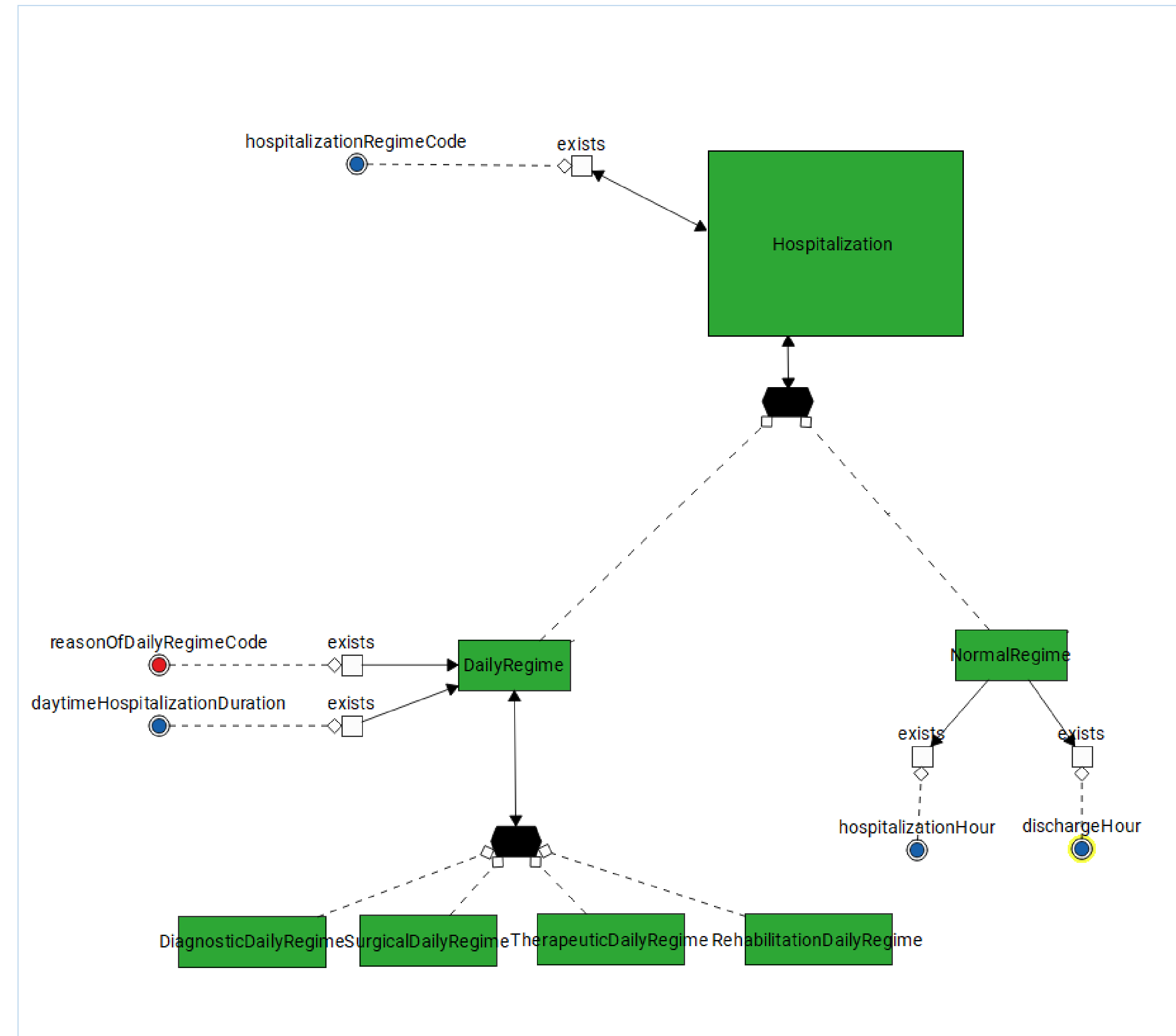


Diagram representing the data flow from regional files to the RIAP database, in the situation of a manual data integration task.



An extract from the graphic representation of some of the concepts and relations described by the RIAP ontology.

## MATERIALS AND METHODS

All the different regional flows of the last five years were mapped to an ontology acting as a common data model by using an automatic data integration methodology, based on the principles of data management in computer-science.

Some ad-hoc developed Python modules and Monolith tool were used to create the logical mappings between the weak-semantic knowledge expressed by the database and the strong-semantic one, modelled by the ontology.

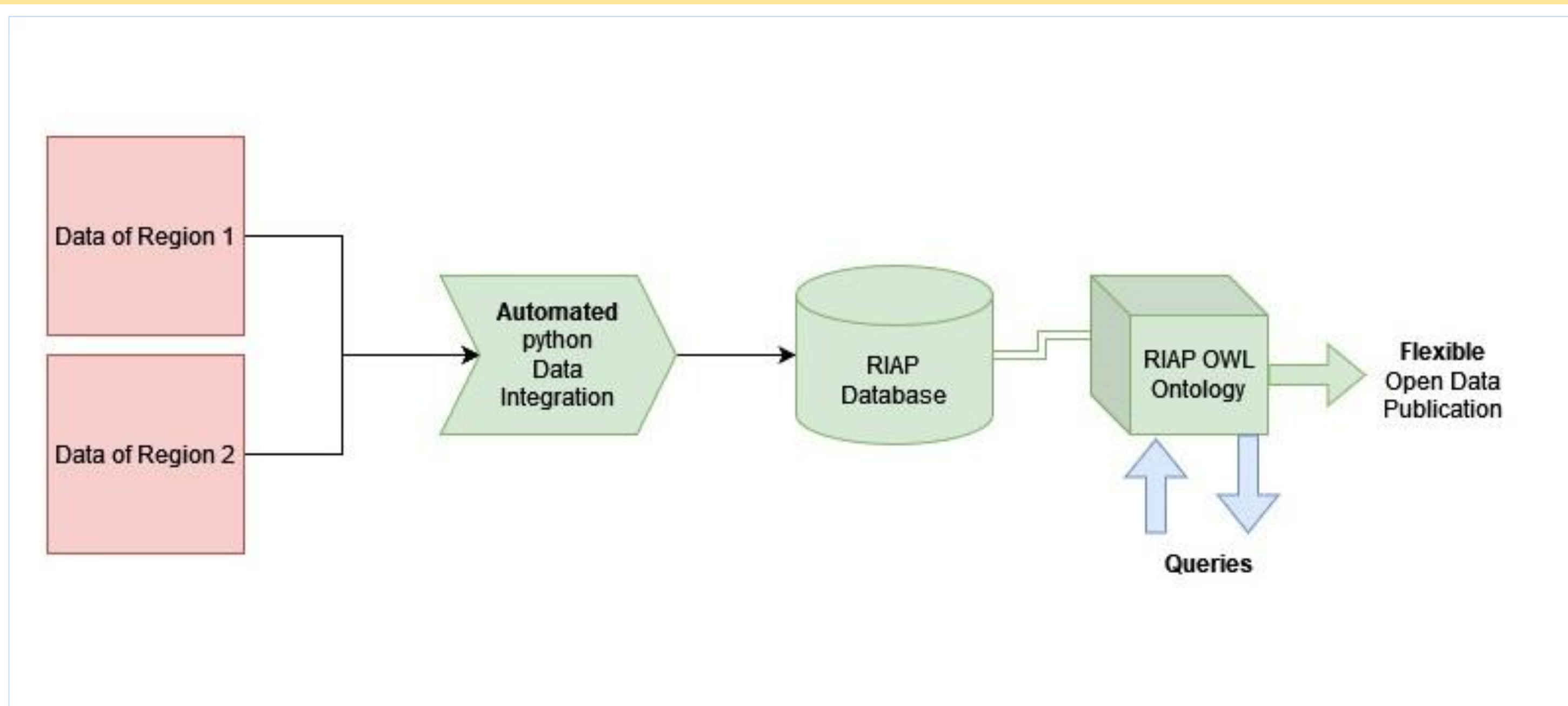


Diagram representing the data flow from regional files to the Ontology, in the scenario of an automated data integration task.

```

1 PREFIX x↔
8
9 select ?i ?td ?cd ?tipd
10 where
11 {
12   ?i a riap:InterventoAnca.
13   ?i riap:haCausaIntervento ?c.
14   ?i riap:haTipoInterventoPrecedente ?tip.
15   ?i riap:haTipoIntervento ?t.
16   ?c riap:descrizioneCausaIntervento ?cd.
17   ?t riap:descrizioneTipoIntervento ?td.
18   ?tip riap:descrizioneTipoInterventoPrecedente ?tipd.
19
20 }
21 order by ?i
  
```

SPARQL query formulation of the task of finding all the arthroplasties, coupled with their causes, type and previous surgery. This kind of query language is useful in the context of SPARQL endpoint, put in place to enable the publication of flexible open data.

## RESULTS

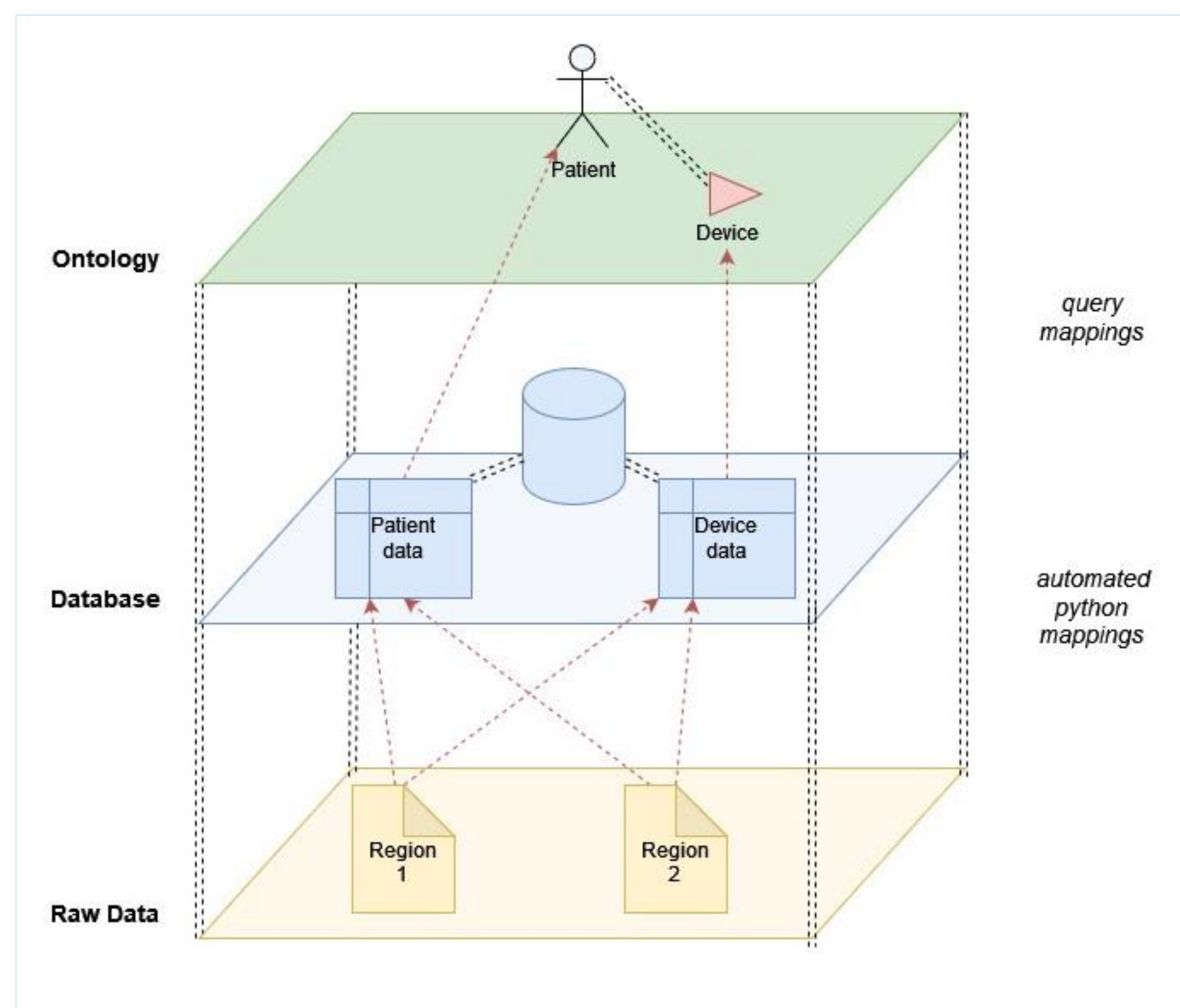
The RIAP database was linked to the modeled ontology to develop a system able to provide user-specific open data by using SPARQL, the standard language used to obtain strong-semantic answers for queries to the ontology.

## DISCUSSION

As a first benefit, the data integration and standardization phase was consistently sped up by the automation process.

Also, thanks to the automation, data quality is expected to improve by reducing the number of human errors involved in the related steps. Furthermore, the ontology layer modeled on top of the RIAP database, paves the way to the future flexible open data publication.

Finally, the described methodology can be also applied to define a common standard among several national registries.



The three stages of an ontology-based data management process: from the raw data, to the ontology.