An Event-Driven Architecture for Biomedical Data Integration and Interoperability

Pedro Lopes and José Luís Oliveira

Universidade de Aveiro, Campus Universitário de Santiago 3810-193 Aveiro, Portugal
{pedrolopes,jlo}@ua.pt

Abstract. Connecting data and services is nowadays an essential skill for life sciences researchers. Handling data from unrelated sources using problem-specific software or labor-intensive tools are common tasks. Despite the latest advances, integration and interoperability developments still involve primitive interactions and manual triggers. On the one hand, available tools cover specific niches, ignoring more generic problems. On the other hand, overly complex tools with excessive features dominate the market. In this proposal we introduce a cloud-based architecture to simplify real-time integration and interoperability for biomedical data. We support our strategy in an event-driven service-oriented architecture, where new data are pushed from any content source, through an intelligent proxy, for delivery to heterogeneous endpoints. This enables a passive integration approach, providing developers with a streamlined solution for deploying integrative biomedical applications.

Keywords: Distributed/Internet based software engineering tools and techniques, interoperability, data translation, data warehouse and repository, bioinformatics.

1 Introduction

The biomedical knowledge landscape, with ever growing datasets, increasingly demands for hardware and software progress [1]. The true value behind existing raw data lies in its adequate exploitation. However, data by itself are no longer valuable if there are no additional services for their better exploration [2].

Life sciences datasets are challenging: the structure is heterogeneous, they originate from distinct places, they adopt multiple standards, the content is stored in different formats and their meaning is always changing [3]. From next generation sequencing hardware [4] to the growing availability of biomedical sensors, exploring these data is an on-going endeavour [5]. Whether through the manual aggregation of data, or the use of specific software, retrieving and integrating data is an everyday requirement in life sciences’ research work.

Recently, biomedical data integration and interoperability is focused on cloud-based service-oriented architectures strategies [6-8]. Cloud-based approaches are useful where the computational requirements are the main bottleneck. For instance, executing intensive analysis algorithms on next-generation sequencing data. Furthermore, using
cloud-based technologies also simplifies common researchers’ tasks, as they do not need to install complex application stacks to perform relatively simple integration operations.

In addition to cloud-based operations, workflow management tools are also responsible for notable advances in biomedical data interoperability. The ability to create comprehensive workflows eased researchers’ work [9]. Nowadays, connecting multiple services and data sources is a trivial task. Yet, tools such as Yabi [10], Galaxy or Taverna [11] lack automation strategies.

Although the adoption of cloud-based and workflow management tools is widespread, there is room for improving their operability, namely through the inclusion of event-driven strategies. These will promote process automation and reduce the burden on researchers and developers alike. Hence, in this work we introduce an architecture to streamline data integration and interoperability. Our objective is to enable the automated event-driven analysis of data, empowering the creation of state-of-the-art integration and interoperability applications.

Traditional integration approaches, in use by warehouses, rely on batch, off-line Extract-Transform-Load (ETL) processes. These are manually triggered on regular intervals of downtime, which can range from weeks to months up to years. However, in the life sciences domain, the demand for fresh data cannot be ignored. Hence, we need to deploy new strategies that are dynamic [12], reactive [13] and event-driven [14, 15]. Modern platforms must act intelligently, i.e., in real-time and autonomously, to new events in integration ecosystems.

Our approach starts with an intelligent event handler, to where any content owner or producer can push newly generated data. When these events occur, the system initiates an Extract-Transform-Load (ETL) process, analysing received data and, when data are new, delivering it to a preconfigured set of endpoints. The delivery process enables interacting with files, databases, emails or URLs. As a result, there are endless combinations for customisable integration tasks, connecting events’ data with actions configured in templates. Among others, the framework empowers live data integration and heterogeneous many-to-many interoperability.

This architecture is targeted at bioinformatics developers. The goal is to enable the creation of new client applications on top of a barebones framework. These cloud-based applications can feature event-driven, publish/subscribe, Integration-as-a-Service architectures.

Biomedical data integration and interoperability is the cornerstone of modern life sciences research projects. In this proposal, we bring forward an innovative architecture to quickly deploy automated and real-time integration workflows.

2 Methods

2.1 Background

This proposal introduces an architecture that can act as the foundation for distinct systems in event-driven environments. Traditional service-oriented architectures