Towards Ease of Building Legos in Assessing eHealth Language Technologies

A RESTful Laboratory for Data and Software

Hanna Suominen[1, 2], Karl Kreiner[3], Mike Wu[1], Leif Hanlen[1, 2]

[1] NICTA, National ICT Australia, Locked Bag 9013, 1435 Alexandria, NSW, Australia
[2] The Australian National University, 0200 Canberra, ACT, Australia
[3] AIT, Austrian Institute of Technology GmbH, Reininghausstraße 13/1, 8020 Graz, Austria
hanna.suominen@nicta.com.au, karl.kreiner@ait.ac.at,
mike.wu@nicta.com.au, leif.hanlen@nicta.com.au

Abstract. More and more scientific literature, care guidelines, health records, social media, and other textual eHealth information are electronically available. Language technologies provide a way to analyse these documents for the benefit of both individuals and populations.

In order to catalyse the development of eHealth language technologies, we propose a virtual laboratory with a standardised platform for easy building and assessment of the systems from the “lego” bricks of shared data, resources, and software. Our aim is to address specific needs in eHealth: governance and sharing of private data; provenance and sharing of resources and software; systematic benchmarking and quality control of systems and their components; and collaboration of eHealth language technology developers and users across healthcare services, academia, industry, and government.

The Epicure virtual laboratory is intended to be used for software and resource evaluation and development as well as for data analysis if data subjects’ privacy is ensured. Epicure is a meta-framework in the sense of abstracting over existing frameworks. Its five roles for clients are data or resource provider, application assembler, application user, software developer, and system administrator.

We have implemented Epicure based on publicly available software. Its control layer is a Glassfish JavaEE server, providing a RESTful (REpresentational State Transfer) application programming interface; web interface for accessing and installing third-party platforms; and easy operation via standard web commands. After proper user authentication and authorisation of incoming requests, it builds applications, analyses data and assesses outcomes by orchestrating storage and execution layers. The storage layer of Epicure uses a CouchDB-based repository for centralised storage of data, resources, and software. It enables controlling document access on the level of documents; tracking all changes; recording these revisions; storing all analysis outcomes; and associating the outcomes with the data, resources and software used in their generation. The execution layer of Epicure provides a runtime environment for executing data analysis tasks and installing third party platforms. It invokes tools as simple commands. A tool must be specify its input format, output formats, parameters, and their possible values as a file and be executable on a command line. Tools do not need to be installed within Epicure itself but instead be accessed via a network interface and wrapper, which provides access from Epicure to this remote service.

Keywords: Evaluation; Health Information Technology; Natural Language Processing; Software Design

1 Introduction

More and more scientific literature, care guidelines, health records, social media, and other *textual eHealth information* is electronically available.[1] *Language technologies* (LTs) provide a way to analyse these documents for the benefit of both individuals and populations. In order to catalyse the development of eHealth LTs, we propose a *virtual laboratory* with a *standardised platform* for easy building and assessment of the systems from the “lego” bricks of shared *data* (e.g., clinical text and annotations), *resources* (e.g., medical dictionaries and data standards), and *software* (e.g., processing, evaluation, and visualisation algorithms).

2 Materials and Methods

The *Epicure* virtual laboratory is a *meta-framework* in the sense of abstracting over existing frameworks (e.g., SNOMED CT Systematized Nomenclature of Medicine Clinical Terms for healthcare terminologies, HL7 Health Level Seven International for interoperability of health information technology, UIMA Unstructured Information Management Architecture for LTs, and WEKA Waikato Environment for Knowledge Analysis). It is named it after an ancient Greek philosophy, which aims to attain a happy, tranquil and self-sufficient life surrounded by friends; similarly, our aim is collaborative building and systematic assessment of eHealth LTs by easy connectivity of data providing, resource and software development, and end use.

Epicure has five roles for *clients*: data or resource provider, application assembler (i.e., build applications from software, resource and data bricks), application user, software developer, and system administrator. It is intended to be used for software and resource evaluation and development as well as for data analysis if data subjects’ privacy is ensured either by limiting data access or conducting appropriate de-identification procedures. *Processing examples* include:

1. choosing a textual dataset and its annotation with respect to a given classification task from the repository, applying a given medical dictionary to reduce data sparseness by synonym and hypernym mappings, training a given classification algorithm to perform the task automatically, and evaluating this classifier by selecting an evaluation method and measures, and
2. submitting a new classification algorithm; choosing the evaluation methods and methods; selecting a dataset to be used in the evaluation and classification algorithms to be compared against; and evaluating the quality of the submitted algorithm via these comparisons.

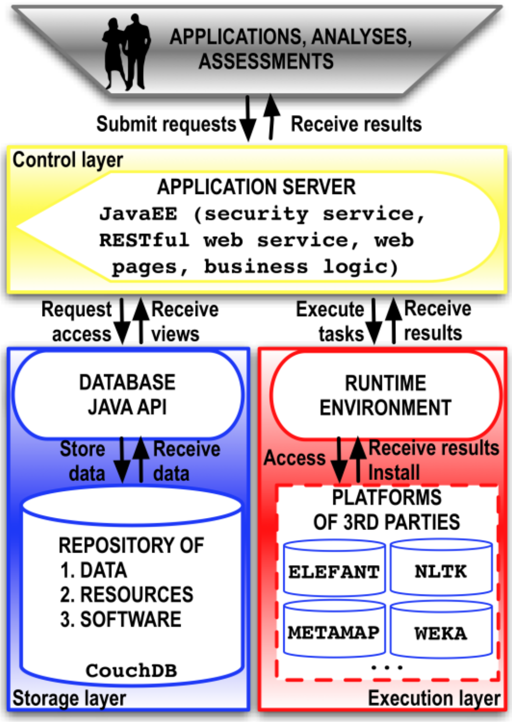

Fig. 1 Epicure meta-framework

Epicure is *implemented* based on publicly available software. REpresentational State Transfer (REST) has been chosen because the same design principles have enabled the success of the World Wide Web.[2] Epicure's main components are (Figure 1):

Control layer is the main communication hub for client interaction. It is a Glassfish JavaEE server, providing a RESTful API (Application Programming Interface); web interface for accessing and installing third-party platforms; and easy operation via the Hyper Text Transfer Protocol (HTTP) commands of GET, POST, PUT, and DELETE for retrieving, creating, replacing, and removing contents. After proper user authentication and authorisation of incoming requests, it builds applications, analyses data and assesses outcomes by orchestrating storage and execution layers. Because many analysis and assessment request take a long time to complete, the control layer must support processing requests asynchronously.

Storage layer: A repository, based on CouchDB, is used for centralised storage of data, resources, and software. It treats data, resources, and software equally as documents which enables controlling document access on the level of documents; tracking all changes; recording these revisions; storing all analysis outcomes; and associating the outcomes with the data, resources and software used in their generation. All data to be analysed on Epicure needs to be stored in this repository.

Execution layer: A runtime environment is provided for executing data analysis tasks and installing third party platforms for machine learning, natural language processing (NLP), data formatting, and information visualisation. It invokes tools as simple commands. A tool must be specify its meta information (i.e., input format, output formats, parameters and their possible values) as an XML (Extensible Markup Language) file and be executable on a command line (and not invoke a graphical user-interface). Tools do not need to be installed within Epicure itself but instead be accessed via a network interface and wrapper, which provides access from Epicure to this remote service. Our plan is to improve this layer by a Hadoop-based implementation, which provides a capability to perform parallel computing and integrate software bricks in all programming languages.

3 Results and Discussion

We have designed and implemented the Epicure virtual laboratory for integrating and sharing data, resources, and software related to building and assessing eHealth LTs. It addresses specific needs in eHealth: governance and sharing of private data; provenance and sharing of resources and software; systematic benchmarking and quality control of systems and their components; and collaboration of eHealth LT developers and users across healthcare services, academia, industry, and government. We have showcased Epicure by building an application for in-hospital surveillance of fungal infections based on patient record texts, expert annotations and authentic outcomes of healthcare.[3] Probably the most similar approach is the iDASH NLP Ecosystem (`nlp-ecosystem.sdsc.edu`), established in 2011. This is a virtual machine with a suite of installed eHealth solutions and capability to download the suite. To differentiate this work from Epicure, the catalyst effect of community collaboration and ease of building systems from standardised bricks of data, resources, and software may be challenged without REST.

4 Conclusion

Epicure promotes reproducibility and comparability of results; availability of data, resources, software and applications; and renewal of science and healthcare practice.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. This work took place during Karl Kreiner's research visit at NICTA.

References

1. Suominen H and Salakoski T. Supporting communication and decision making in Finnish intensive care with language technology. *Journal of Healthcare Engineering* 2010 1(4), 595-614.
2. Fielding RT (2000). *Architectural Styles and the Design of Network-based Software Architectures*; bit.ly/3jFBnu.
3. Martinez D, Suominen H, Ananda-Rajah M, and Cavedon L. Biosurveillance for invasive fungal infections via text mining. *Proceedings of the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis (CLEFeHealth2012),* Rome, Italy, 17–20 September 2012.