An Ontology and a REST API for Sequence Based Microbial Typing Data

Article · January 2010

DOI: 10.1007/978-3-642-28062-7_3

7 authors, including:

Mário Ramirez
University of Lisbon
261 PUBLICATIONS **4,678** CITATIONS

SEE PROFILE

João A Carriço
University of Lisbon
120 PUBLICATIONS **3,268** CITATIONS

SEE PROFILE

An Ontology and a REST API for Sequence Based Microbial Typing Data

João Almeida[1⋆], João Tiple[1⋆], Mário Ramirez[2], José Melo-Cristino[2], Cátia Vaz[1,3], Alexandre P. Francisco[3,4], and João A. Carriço[2]

[1] DEETC, ISEL, Poly Inst of Lisbon
[2] IM / IMM, FM, Univ of Lisbon
[3] INESC-ID Lisbon
[4] CSE Dept, IST, Tech Univ of Lisbon

Abstract. In the Microbial typing field, the need to have a common understanding of the concepts described and the ability to share results within the community is an increasingly important requisite for the continued development of portable and accurate sequence-based typing methods. These methods are used for bacterial strain identification and are fundamental tools in Clinical Microbiology and Bacterial Population Genetics studies. In this article we propose an ontology designed for the microbial typing field, focusing on the widely used Multi Locus Sequence Typing methodology, and a RESTful API for accessing information systems based on the proposed ontology. This constitutes an important first step to accurately describe, analyze, curate, and manage information for microbial typing methodologies based on sequence based typing methodologies, and allows for the future integration with data analysis Web services.

Keywords: ontology, knowledge representation, data as a service, microbial typing methods

1 Introduction

Microbial typing methods are fundamental tools for the epidemiological studies of bacterial populations [7]. These techniques allow the characterization of bacteria at the strain level providing researchers important information for the surveillance of infectious diseases, outbreak investigation and control, pathogenesis and natural history of an infection and bacterial population genetics. These areas of research have a direct impact in several human health issues, such as in the development of drug therapies and vaccines [1], with the concomitant social and economical repercussions.

With the decreasing cost and increasing availability of DNA sequencing technologies, sequence based typing methods are being preferred over traditional molecular methodologies. The large appeal of sequence-based typing methods

⋆ These authors contributed equally to this work.

is the ability to confidently share their results due to their reproducibility and portability, allowing for a global view and immediate comparison of microbial strains, from clinical and research settings all over the world. Several online microbial typing databases have been made available for different methods. The most successful examples are the Multi-Locus Sequence Typing (MLST) [6] databases for a multitude of bacterial species [10,12,8], *emm* typing database for *Streptococcus pyogenes* [14] and *spa* typing for *Staphylococcus aureus* [13].

However, these efforts are not standardized for data sharing, suffering from several caveats, being the most notable the lack of interfaces for automatic querying and running analysis tools. The automatic integration of data from the different databases is also hindered due to the lack of common identifiers among different databases. Moreover, the absence of an automatic validation of the new data in the submission process is leading to an increase of incomplete and unreliable data in the majority of these databases, seriously hampering the promised advantages of methodological accuracy and portability of results between laboratories. This is even more significant with the rise of new Single Nucleotide Polymorphism (SNP) typing techniques based upon the Next Generation Sequencing (NGS) [4] methods. The validity of this new high-throughput technology can be seriously hampered if the complete data analysis pipeline cannot be fully described in public databases, in order for the results to be reproducible. Also, the ability to integrate information from several well established typing methodologies will be paramount for the validation and development of the more informative whole genome approaches [5,3] based on these NGS methods for the bacterial typing field.

In a largely descriptive science such as Microbiology, the need to have a common understanding of the concepts described is fundamental for continued development of sequence-based typing methods. Therefore, the definition of an ontology that can validate and aggregate the knowledge of the existing microbial typing methods, is a necessary prerequisite for data integration in this field. In order to solve those problems, we present in this paper the design and implementation of an ontology created for the microbial typing field and an Application Programming Interface (API) to an information system using the concepts of the REST (Representational State Transfer) paradigm [2]. The proof-of-concept prototype of the proposed framework, focusing on the well established MLST methodology, is available at `http://rest.phyloviz.net`.

The ability to accurately describe the relationships between typing methods through the use of an ontology and to offer REST services to analyze, curate, and manage the information will facilitate the implementation of information systems capable of coping with the heterogeneous types of data existing in the field, including the re-usage of legacy data formats and methods.

This paper is organized as follows. Section 2 describes the proposed ontology, TypOn. Section 3 presents a REST API suitable for managing microbial typing data. Section 4 briefly details the RESTful Web services prototype implementation. Finally, Section 5 provides some final remarks and future work directions.

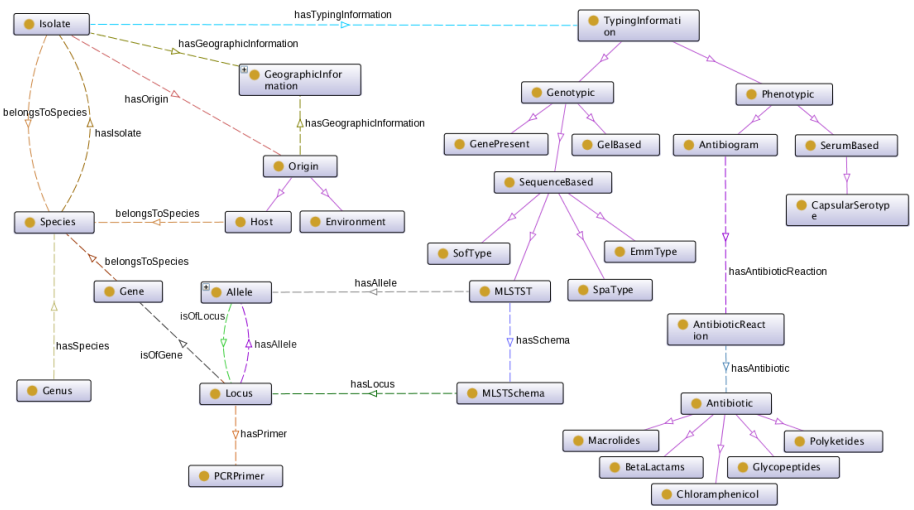

Fig. 1. TypOn, an ontology for microbial typing data. Dashed lines represent object properties and solid lines represent subclass relations, e.g., *Host is-a Origin*.

2 TypOn – Typing Ontology

An ontology should make available both the vocabulary and the semantic rules required to proper represent knowledge of a given domain. In this section we provide an ontology suitable to describe knowledge in the microbiology typing methods domain, TypOn, depicted in Fig. 1. This ontology was developed and improved based on comments by domain experts and it constitutes a first proposal, that can be expanded and adapted as new typing methods are developed and already existing ones are updated. The ontology was developed with the help of the Protégé editor [11] and is available at `http://www.phyloviz.net/typon`.

The main aim of bacterial typing methods is the characterization of bacterial populations, where each sampled microorganism becomes an isolate, referring to the process of isolating it from the bacterial population. Thus, *Isolate* is a main concept for TyPon and it is characterized by several properties. An isolate belongs to a *Species*, property *belongsToSpecies*, which makes part of a *Genus*, property *hasSpecies*. The property *belongsToSpecies* has the property *hasIsolate* as its inverse. Moreover, for each *Isolate*, we know its *Origin*, either *Host* or *Environment*, its *GeographicInformation* and its *TypingInformation*. Note that a *Host* belongs also to a *Species* and that both *Host* and *Environment* may also have *GeographicInformation*. Although properties *hasGeographicInformation* and *hasOrigin* have usually cardinality at most one for each *Isolate*, the property *hasTypingInformation* has usually cardinality higher than one for each *Isolate*. For instance, an *Isolate* usually has available information for several typing methodologies such as MLST, antibiograms, etc. In this context, it is important to note that *TypingInformation* is the root of a class hierarchy which is

extensible and that defines several typing methods (see Fig. 1). In particular, we are able to distinguish different categories of typing methods, *e.g.*, the ontology allow us to infer that *MSLTST* is a *Genotypic* technique and that, in contrast, *Antibiogram* is a *Phenotypic* technique.

As mentioned before, the current version of TyPon focus on MLST concepts, since it is the most widely used sequence based typing technique. In this context, we note in particular the concepts *Locus*, *Allele*, *MLSTSchema* and *MLSTST*. In MLST we can have several typing schemas described by a set of loci, each one being part of a sequence of an housekeeping gene. Such schemas are represented through the class *MLSTSchema*, which has the property *hasLocus*. Then, each *Isolate* may have associated one or more typing informations, obtained with different schemas, i.e., *MLSTST* instances, known as sequence types characterized by the observed alleles for each locus. Therefore, in our ontology, we associate to each *MLSTST* both a schema and the observed alleles through properties *hasSchema* and *hasAllele*, respectively. Notice also that *hasAllele* is a property shared by *MLSTST* and *Locus* classes and, thus, it does not have *isLocus* property as its inverse. It is also interesting to note that, by knowing only the *Locus*, it is possible to be aware of the *Species* that it belongs to, using the *isOfGene* and *belongsToSpecies* properties. The property *belongsToSpecies* is also an example of a property which has more than one class as domain.

We have also detailed the *Antibiogram* typing information technique in the current version. Namely, we have represented each *Antibiotic* as a concept, allowing the addition of new antibiotics as needed. The reaction of a given antibiotic is also represented as a concept, *AntibioticReaction*, allowing that each *Antibiogram* may have associated one or more antibiotic reactions, depending on the number of used antibiotics. These relations are given through the object properties *hasAntibioticReaction* and *hasAntibiotic*, respectively.

Additional information for each class, such as *id* and *other name*, are described through data properties. For instance, the class *GeographicInformation* has data properties such as *Country* and *Region*. The class *Isolate* has data properties such as *Strain* and *Year*.

3 RESTful Web services

A second contribution of our work is a RESTful API for making available microbial typing data represented through the above ontology. A Web services framework is under development, making use of the Jena Semantic Web Framework [9] and other standard Java technologies for developing Web services. The set of endpoints that were defined for retrieving microbial typing data include:

```
/services/typingmethods
/services/{typingmethod}
/services/{typingmethod}/genus
/services/{typingmethod}/{genusid}
/services/{typingmethod}/{genusid}/species
/services/{typingmethod}/{genusid}/{speciesid}
```

`/services/{typingmethod}/{genusid}/{speciesid}/isolates`
`/services/{typingmethod}/{genusid}/{speciesid}/{isolateid}`
`/services/{typingmethod}/{genusid}/{speciesid}/sts`
`/services/{typingmethod}/{genusid}/{speciesid}/sts/{stid}`
`/services/{typingmethod}/{genusid}/{speciesid}/sts/{stid}/isolates`
`/services/{typingmethod}/{genusid}/{speciesid}/loci`
`/services/{typingmethod}/{genusid}/{speciesid}/loci/{locus}`
`/services/{typingmethod}/{genusid}/{speciesid}/loci/{locus}/{id}`

The URI parameters, *i.e.*, the text inside {}'s, represent specific identifiers. For instance, {`typingmethod`}, {`genusid`} and {`speciesid`} should be parametrized with the name of the typing method (*e.g.* MLST), the name of the genus (*e.g. Streptococus*) and the name of the species (*e.g. pneumoniae*), respectively.

Each endpoint with {}'s at the end refers to a resource identified by a given id or unique label. As an example, with the endpoint

`/services/{typingmethod}/{genusid}/{speciesid}/sts/{stid}`

we may obtain the information of a specific sequence type. Moreover, with these kind of endpoints it also possible to replace their information, using the POST method. The other endpoints retrieve all individuals of a respective class. For instance, the endpoint

`/services/{typingmethod}/{genusid}/{speciesid}/sts`

retrieves all existing MLST sequence types in the database for the specified parameters {`typingmethod`}, {`genusid`} and {`speciesid`}. We can also add more individuals with these kind of endpoints, using the PUT method. However, data deletion is only possible through the endpoints

`/services/{typingmethod}/{genusid}/{speciesid}`
`/services/{typingmethod}/{genusid}/{speciesid}/{isolateid}`

by using the DELETE method.

All endpoints return either `text/html` or `application/json`. There is also available a SPARQL endpoint and an authenticated endpoint to retrieve and submit data represented as `rdf/xml`. A more comprehensive description for the MLST data related endpoints is available at `http://rest.phyloviz.net`.

4 Implementation

A prototype Web client that makes use of the RESTful API and that allows users to explore and query data for some of the MLST public datasets, is also available at `http://rest.phyloviz.net/webui/`. In this prototype it is possible to query by MLST schema, *MLSTSchema*, by the id of the sequence type, *MLSTST*, and by locus, *Locus*. Also, the MLST schema and alleles can be downloaded in more than one format. A graphical visualization of isolate statistics is also available in this prototype.

Our implementation makes use of the Jena Semantic Web Framework [9] and other standard Java technologies for developing Web services. Jena provides an API to deal with RDF data, namely a SPARQL processor for querying RDF data. In our implementation, both TypOn and all typing data are stored as RDF statements on a triple store. We are currently using the TDB triple store, a component of Jena for RDF storage and query. Although the Jena framework can use several reasoners, including OWL-DL reasoners, we are using the internal RDFS reasoner for validation purposes

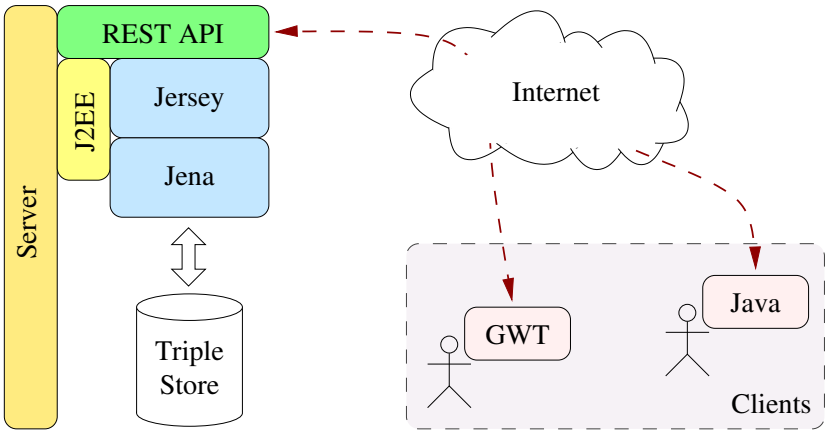

Fig. 2. Architecture of the Web service prototype. A REST API implemented over the Jersey framework, is made available, where data is accessed through the Jena framework. On the client side, we have implemented a Java REST client library and a Web application implemented over the Google Web Toolkit (GWT).

only. Nevertheless, given Jena flexibility, we can easily process our repository of statements through a more powerful reasoner, and insert inferred and relevant statements back to our repository. This is particularly useful whenever we update the ontology with new or equivalent concepts and properties, or when we want to index frequent SPARQL queries, in order to improve their speed. Moreover, under the open world assumption, with data distributed over several repositories, one may need to crawl and index several repositories, possibly instances of our Web service implementation, before proceed with reasoning and inference.

The REST API made available uses the Jersey implementation of JAX-RS (JSR 311), a Java API for RESTful Web services that provides support in creating Web services according to the REST architectural style. This implementation is an official part of Java EE 5 and it facilitates the implementations of RESTful Web services through the usage of annotations, simplifying the development and deployment of Web service clients and endpoints.

In the current implementation, any user can query the repository and only authenticated users can insert, update or delete data. A more refined authorization model is under development.

5 Final remarks

The proposed ontology provides the basic concepts needed to establish the semantic relationships of the different sequence-based typing methodologies, and it is designed to allow further expansion. It should be easily expanded to encompass the newer NGS SNP typing techniques that are appearing in the microbial typing field, while providing a consistent link with legacy techniques and other databases. This Semantic Web approach for sharing microbial typing data also allows for local databases from differ-

ent institutes and different methods to be connected through the use of specific REST endpoints.

Moreover, the proposed REST interface and ontology facilitates the decoupling between the information system and its possible client technologies, allowing the sharing of data in human- and machine-readable formats. This approach allows the design of novel interfaces between different databases and data analysis softwares, through the use of Web services mashups.

An immediate practical use of the framework is to provide the microbiology researchers with a quick and effective way to share data on new methods being developed based on sequence typing methods, since the creation of a new typing schema and adding its concepts on the ontology is straightforward. The information available for isolates typed using a new typing schema can then be parsed to RDF statements and uploaded to a server authenticated SPARQL endpoint and, then, a new database is automatically accessible. The GWT Web client provides then to the end-users a friendly interface for data access for querying and submitting new data.

Future work will focus on expanding the ontology and creating Web services to perform automated curation of data directly from sequencer files, in order to speed up the curation process, and ensure better quality and reproducibility of data in the field of microbial typing.

Acknowledgments. The work presented in this paper made use of data available at MLST.net [10], PubMLST [12] and Institut Pasteur MLST Databases [8].

References

1. Aguiar, S., Serrano, I., Pinto, F., Melo-Cristino, J., Ramirez, M.: Changes in Streptococcus pneumoniae serotypes causing invasive disease with non-universal vaccination coverage of the seven-valent conjugate vaccine. Clinical Microbiology and Infection 14(9), 835–843 (2008)
2. Fielding, R.: Architectural styles and the design of network-based software architectures. Ph.D. thesis, Citeseer (2000)
3. Harris, S., Feil, E., Holden, M., Quail, M., Nickerson, E., Chantratita, N., Gardete, S., Tavares, A., Day, N., Lindsay, J., et al.: Evolution of MRSA during hospital transmission and intercontinental spread. Science 327(5964), 469 (2010)
4. MacLean, D., Jones, J., Studholme, D.: Application of'next-generation'sequencing technologies to microbial genetics. Nature Reviews Microbiology 7(4), 287–296 (2009)
5. Mwangi, M., Wu, S., Zhou, Y., Sieradzki, K., De Lencastre, H., Richardson, P., Bruce, D., Rubin, E., Myers, E., Siggia, E., et al.: Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. Proceedings of the National Academy of Sciences 104(22), 9451 (2007)
6. Spratt, B.: Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. Current opinion in microbiology 2(3), 312–316 (1999)
7. Van Belkum, A., Struelens, M., De Visser, A., Verbrugh, H., Tibayrenc, M.: Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clinical microbiology reviews 14(3), 547 (2001)
8. Institut Pasteur MLST Databases. `http://www.pasteur.fr/mlst/`, Pasteur Institute

9. Jena A Semantic Web Framework for Java. http://jena.sourceforge.net/, HP and Others
10. MLST: Multi Locus Sequence Typing. http://www.mlst.net, Imperial College of London
11. The Protégé Ontology Editor and Knowledge Acquisition System. http://protege.stanford.edu, Stanford Center for Biomedical Informatics Research
12. PubMLST. http://pubmlst.org/, University of Oxford (UK)
13. Ridom SpaServer. http://www.spaserver.ridom.de/, Ridom bioinformatics
14. *Streptococcus pyogenes emm* sequence database. http://www.cdc.gov/ncidod/biotech/strep/M-ProteinGene_typing.htm, CDC