A comprehensive microbial knowledge base to support the development of *in-vitro* diagnostic solutions in infectious diseases

Magali Jaillard[1], Stéphane Schicklin[1], Audrey Larue-Triolet[1], and Jean-Baptiste Veyrieras[1]

Data & Knowledge Lab, Technology Research Department, bioMérieux SA, France
magali.dancette@biomerieux.com

1 Background

Research and development of innovative *in-vitro* diagnostic (IVD) solutions in infectious diseases require to federate up-to-date knowledge from several fields known for their complexity and their constant evolution: medical practices, microbiology and system and software engineering [6, 11, 12].
To tackle the inherent complexity of such multidisciplinary R&D projects, modern information technologies now offer powerful environments which can be leveraged to facilitate information sharing between corporate experts. This is key to ensure semantic alignments, information retrieval and then to foster decision making within the projects. The advent of almost mature semantic technologies together with international standards bring the possibility to create enterprise compliant knowledge bases. The major challenge is then to gather and link all the information from distinct and heterogeneous sources in a frequently updated and fully searchable resource. Ideally, for IVD projects, such a resource would allow for instance to map unmet needs onto current medical practices in infectious diseases, to facilitate comparison of results from different technologies, or to gather and maintain pathogen-related knowledge.
Towards this goal, we benefited from the recent efforts from the biomedical and bioinformatics communities which have been early adopters of the promising web 3.0 functionalities; multiple public data resources have developed and released domain specific ontology models or SPARQL endpoints [5, 8, 14]. Taking advantage of these semantic components we deployed on the company intranet BioPedia, a private collaborative semantic web platform carrying a cross domain knowledge base dedicated to human pathogens. The knowledge is stored on a triplestore while a wiki-based interface allows to create powerful faceted queries.

2 Methods

The current architecture of Biopedia is based on a central triplestore interfaced with sparql 1.1 compliant 4store [10] endpoint providing full sparql query and sparul update functionalities. The display and query of the triplestore content relie on several semantic wikis covering specific domains (Figure 1 A.). A benchmark of semantic solutions led our choice to MediaWiki (MW) [4] framework and

Semantic MediaWiki (SMW) [13] extensions which provide an always growing palette of querying tools. The global ontology describes four domains: bacteria and fungi strains (BioSource), taxonomy nomenclature and classification (BioTaxon), determinants and resistance mechanisms (BioGraM) and genomic data (BioSeq) (Figure 1 B.), laying on the following main classes:

- Strain: variant of a microorganism; distinct strains differ by their genomes
- Taxon: unit of close strain group, associated to a label (such as *Escherichia coli*) and a rank (for instance *species*)
- Genome: entire genetic information as chromosome and plasmid sequences
- Locus: sub-sequence of a genome annotated for its functionality
- Resistance determinant: a mutation, single nucleotide polymorphism, gene, or gene product that confers antibiotic resistance
- Antimicrobial: agent that kills microorganisms or inhibits their growth

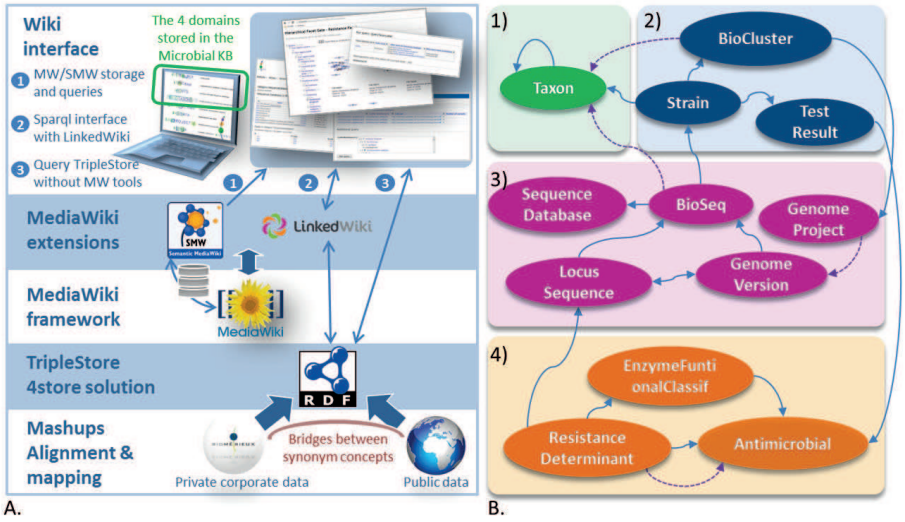

Fig. 1. Overview of the Microbial Knowledge Base structure.
A. Architecture and wiki interface of the solution, mainly based on a 4store endpoint and MediaWiki framework. **B.** Simplified view of the ontology showing the main classes and their relationships for the four domains 1) Biotaxon, 2) BioSource, 3) BioSeq and 4) BioGraM. Dashed lines represent inferred relationships.

The triplestore is populated with mashed up data mapped on the ontology. The mashup of data from heterogeneous sources includes ontology alignments, terms mapping or bridges between synonym concepts from the company and from public sources. BioTaxon domain contains bridges translating corporate identifiers to NCBI [3] taxon identifiers. These taxon identifiers are mapped using their associated taxon labels as there are the most standardized shared data. Indeed, the International Committee on Systematics of Prokaryotes [2] (ICSP) regularly publishes nomenclature rules for microbes used by the scientific community.

A crucial point to populate BioSource domain is to first identify and gather equivalent strains, *i.e.* strains issued from one unique sample and multiplied by creating subcultures. To do so we set up a clustering process using internal and external strain cross-references as edges to deduce connected components with the igraph R library [7]. We selected 75 strain reference collections, and collected strain identifiers belonging to them through StrainInfo [8], the PathoSystems Resource Integration Center (PATRIC) [9] and internal databases. Each BioCluster thus obtained was then connected to a Taxon instance. However a cross validation was necessary to highlight discrepancies: within one cluster, all strains should be tagged with the same taxon identifier. This is not the case when there are annotation or strain identification errors.
A mapping between Strain and Genome was built in order to federate public genome data from PATRIC and from our internal genome database and thus populate the BioSeq domain. Genome sequences can be processed to provide annotation that can be used as one source to populate the Locus class. Here, Loci that are registered as Resistance determinants can give a very valuable information about the strain ability to resist to antimicrobials. BioGraM is the alignment result between our corporate master data knowledge base and the Comprehensive Antibiotic Resistance Database (CARD) [1] (mainly Resistance Determinants and Antimicrobials classes).

3 Results

The current triplestore contains more than 14 million triples linking the four domains of BioPedia and allowing to infer new knowledge (Figure 2). The bridge between the taxonomy nomenclatures of NCBI and our corporate reference taxonomy was built using the taxon translator tools we developed. However 10% of the corporate labels could not be mapped on NCBI taxon labels (Figure 2 A. 3)). This is partially due to a shift of nomenclature versions between sources. Indeed, because labels are not standardized, NCBI can still uses *Fluoribacter bozemanae* when ICSP suggests *Legionella bozemanae*. For these labels, the bridge is manually completed by our expert taxonomy curator.
As shown on Figure 2 A. 1), public strains were much more reduced when gathered into clusters as there are very connected data while private strain have fewer strain cross-references to clusterize. Among the 114,383 strain clusters in which at least one strain belongs to our corporate collection, 8% were allowed to link much more metadata such as public genomes. The validation process also highlighted 2% of clusters whose strains did not share a common taxon identifier. A sparse matrix is then used to help the curator to identify the incriminated vertices.
The integration of this content in the semantic web portal BioPedia provides powerful querying tools such as a hierarchical browser to navigate within the taxonomy classification or faceted searches based on semantic properties. Together with the sparql facilitator provided by the LinkedWiki extension, this allows us bringing a solution for R&D project teams to (i) easily federate all the

available data generated so far for any pathogen stored into the global strain collection or (ii) create reference strain panels based on various criteria depending on targeted diagnostic applications (Figure 2 B.).

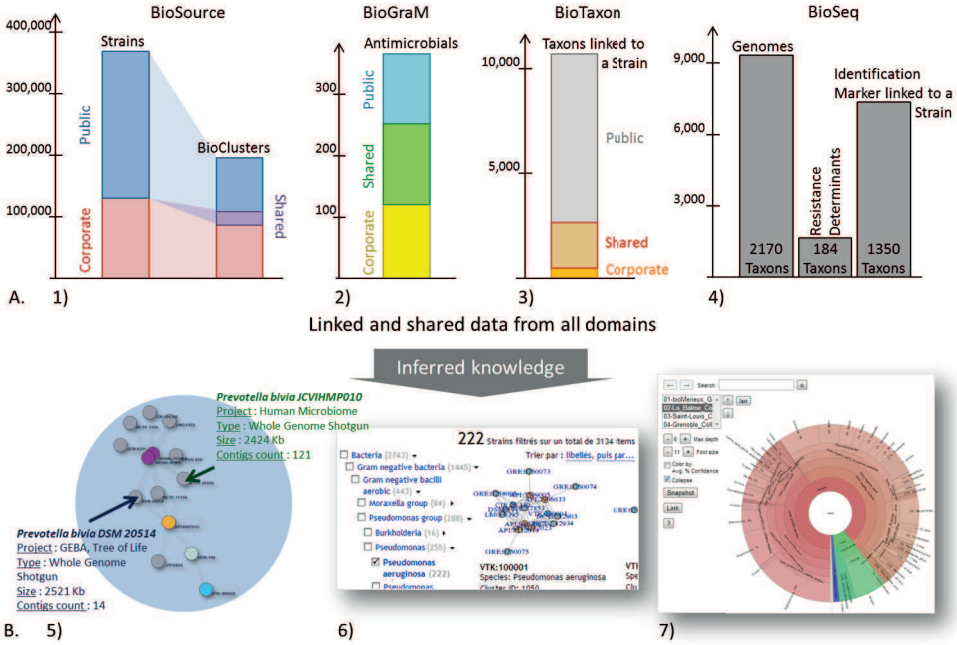

Fig. 2. Mashup results (A.) and examples of the querying capabilities (B.). **1)** Strain and clusters from public and corporate sources **2)** Alignments on antimicrobial terms. **3)** Mapping of taxon labels. **4)** Genome and locus sequence content. **5)** Inferring and chosing the best public genome related to a corporate strain. **6)** Hierarchical gate to explore clusters using the taxonomy. **7)** Global view of a strain panel, clearly showing the taxonomic classification.

4 Discussion

The microbial knowledge base provides global and uniform knowledge of the company strain collection and links it to many infectious diseases oriented public metadata, such as resistance to antimicrobials or genomes. This work gives an enriched overview of this strain collection and connects it to the achievements of the scientific community.

Then, as a side-benefit, linking data from several sources through a semantic store is of great help to improve data quality. Indeed in the mashups, sibling concepts from heterogeneous information streams are blended together and this new closeness drastically highlights the discrepancies. The data curation, even semi-automatic, is time-consuming but mandatory to build a trustworthy reference knowledge base on which powerful queries can be launched and reference datasets can be exported with confidence.

The resulting collaborative semantic web service makes possible to connect heterogeneous data in a corporate way. As the access to data is centralized, it avoids data silo and data tomb often caught out in excel spread-sheets without associated metadata. Moreover the collaborative aspect of this system encourages scientific experts to complete missing information that are then validated by a moderator, thus participating to the enrichment and quality increase of the knowledge base.

References

1. Comprehensive antibiotic resistance database, mcmaster university, canada. http://arpcard.mcmaster.ca.
2. International committee on systematics of prokaryotes). http://www.the-icsp.org/.
3. A. Acland, R. Agarwala, T. Barrett, J. Beck, D. A. Benson, C. Bollin, E. Bolton, S. H. Bryant, K. Canese, D. M. Church, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 41(D1):D8–D20, 2013.
4. D. J. Barrett. *MediaWiki*. O'Reilly Media, Inc., 1 edition, 2008.
5. F. o. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, J. Morissette, et al. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5):706–716, 2008.
6. L. Bissonnette and M. G. Bergeron. Next revolution in the molecular theranostics of infectious diseases: microfabricated systems for personalized medicine. *Expert review of molecular diagnostics*, 6(3):433–450, 2006.
7. G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695:38, 2006.
8. P. Dawyndt, M. Vancanneyt, H. De Meyer, and J. Swings. Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. *Knowledge and Data Engineering, IEEE Transactions on*, 17(8):1111–1126, 2005.
9. J. J. Gillespie, A. R. Wattam, S. A. Cammer, J. L. Gabbard, M. P. Shukla, O. Dalay, T. Driscoll, D. Hix, S. P. Mane, C. Mao, et al. Patric: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and immunity*, 79(11):4286–4298, 2011.
10. S. Harris, N. Lamb, and N. Shadbolt. 4store: The design and implementation of a clustered rdf store. In *5th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2009)*, pages 94–109, 2009.
11. M. Ieven, R. Finch, and A. van Belkum. European quality clearance of new microbiological diagnostics. *Clinical Microbiology and Infection*, 19(1):29–38, 2013.
12. T. A. Metcalfe. Development of novel ivd assays: a manufacturer's perspective. *Scandinavian Journal of Clinical & Laboratory Investigation*, 70(S242):23–26, 2010.
13. M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 585–594, New York, NY, USA, 2006. ACM.
14. P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl 2):W541–W545, 2011.