Abstract—In this annotated bibliography, I will present 25 references that can be use in the design and implementation of a digital library aiming at managing the huge amount of data coming from malaria models.

Index Terms—Malaria, Modeling, \LaTeX, Digital library, Bioinformatics, Data warehouse.

I. INTRODUCTION

Modeling Malaria is not only modeling climate, vectors or population, but also leveraging different existing models and storing the data they need to run and the data they produce in a consistent way. Without appropriate data management, it is nearly impossible to query and analyze big sets of cross model data. Consequently it is important to look for an organized way to handle those data and to be able to answer high level questions about the disease.

II. ANNOTATIONS

A. Model papers

To be able to build the infrastructure, we first need to understand how the models work.

[1] Comparing the Effectiveness of Malaria Vector-Control Interventions through a Mathematical Model
This paper explains how mathematical modeling is used to simulate malaria. Furthermore, they give a use case to compare the effectiveness of treated bed nets compared to indoor spraying.

This paper explains how the EMOD Model works and how it is effectively used not only to compare interventions effectiveness but also how it is built to provide a complete modeling framework allowing to discover synergies between interventions.

This paper describes another model called the Imperial model. It will be good to dive into it a little bit deeper to understand how this model is built and how it is used.

[4] Modeling the impact of vector control interventions on Anopheles gambiae population dynamics
This paper focuses on another malaria model. This one adds an interesting precision on the early mosquito lifecycle (larvae stages) by taking into account the larval density. This allows a more precise modeling of the different interventions that impacts directly or indirectly the larval stages.

B. Data management and Digital Library

The data generated by the models needs to be properly stored and managed to ensure reusability and cross model analysis. We often refer to this system as a digital library. Semantic technologies will also play a big role in the data management.

This paper tries to give a solution about the data lifecycle management when it comes to epidemiology simulations. These simulations generate huge amounts of numerical data and require an accurate management that this paper offers to describe.

This paper presents a framework to use statistical data in a semantic way. It will be interesting to exploit this framework to transform model output data into semantic data.

[7] Transforming Statistical Linked Data for Use in OLAP Systems
This paper focuses on the process of using statistical semantic data in a data warehousing way using OLAP. Therefore enabling to issue OLAP operations on linked semantic datasets. We could leverage this system to take advantage of the versatility of semantic dataset while conserving the flexibility of OLAP.

[8] Eco-informatics modeling via semantic inference
This paper presents a whole architecture using semantic technologies to manage huge amount of data and provide easy analysis and data mining mechanisms. We can pick some good ideas there to design and implement our system: they give a good amount of details on the design and implementation of their solution.
The RDF Data Cube Vocabulary
This paper describe a vocabulary used to describe the concept of data cube in a semantic way. It can be very useful to use this kind of vocabulary to organize the semantic data in a data cube way allowing then to easily port RDF data to a Data warehouse.

Telemaco: Context-aware System for Tourism Guiding based on Web 3.0 Technology
Telemaco is a web based solution that leverage semantic technologies to develop a tourism guide. Although the field of this project is absolutely not related to malaria, it presents an interesting web architecture and uses Django, the python web framework that I envision using for this project.

Ontology-based Integration of OLAP and Information Retrieval
Talks about building a scientific portal based on ontologies. The interest in this paper comes from the fact that they describe how to retrieve data coming from different sources: some comes from a traditional datawarehouse and some other from unstructured datasets.

SPARQL Query Language for RDF
This W3C paper describes SPARQL a language used to query RDF datasets. If we are using semantic technologies, we will likely be using RDF and so SPARQL seems the appropriate language to build queries.

The semantic web
This is the original article presenting what is the semantic web. This is where it started and reading it is good to know exactly what the semantic web vision is.

OBO and OWL: Leveraging Semantic Web Technologies for the Life Sciences
Presents a mapping between OBO and OWL to leverage the power of reasoners used with OBO ontologies which is a largely used biomedical ontologies repository.

Biological knowledge management: the emerging role of the Semantic Web technologies
This article give a general overview of the benefits of using semantic technology when it comes to manage biological knowledge. This article is a good source of general concept that can be applied to data management.

Improving life sciences information retrieval using semantic web technology
Storing data is an important part but retrieving them in a meaningful manner is also important. Defining an abstraction for knowledge using RDF and generates significant and correct outputs is what this article is about.

Aggregation of bioinformatics data using Semantic Web technology
This paper describe the use of Seamark Navigation a system allowing you to browse and search through your RDF data. Also the use case about drug discovery can be studied to find interesting leads on how to mine and search RDF data.

YeastHub: a semantic web use case for integrating data in the life sciences domain
This paper is about a practical use case: they describe how they have put together a system allowing extracting data from heterogeneous datasets and loading them in a datawarehouse. Then they built a web collaborative portal allowing querying the data using RDF-based queries.

Applying Semantic Web Services to Bioinformatics: Experiences Gained, Lessons Learnt
This paper focuses on three project using semantic web services in the bioinformatics field. Every one of them is based on semantic technologies and emphasis on interoperability. We can learn from their experience and use some of their reflection path for our project.

Special Issue: Portals for life sciencesProviding intuitive access to bioinformatic tools
This article encompass a lot of issues every science portal will need to overcome and present an overview of mature technologies that can be used in the context of these particular portals.

Data Management and Analysis Architecture for a More Efficient and Productive Bioinformatics Environment
Here is presented a nice overview of an architecture to allow scientists to utilize disparate data in their analysis. This architecture focuses on leveraging current solutions to provide a better way to manage data in the bioinformatics field.

GReIC Data Storage: A Lightweight Disk Storage Management Solution for Bioinformatics “in silico” Experiments
When it comes to data management, we also need to include storage of the data itself. This GReIC seems to be a good distributed solution to store data and metadata.

An Extensible, Scalable Architecture for Managing Bioinformatics Data and Analyses
Again, this paper presents an architecture but this one has the advantage of being actually implemented. Furthermore, their workflow creation module seems very interesting and may be used in our system. The fact that the system encompass all the layers: from the data to the client, is noticeable and a lot of good ideas presented here could be salvage.

A Cloud Framework for Parameter Sweeping Data Mining Applications
Exploring the effects of a slight change in a parameter
space is one of the core components of bioinformatics. Data mining the results of a huge amount of simulations is time and resource consuming. This paper addresses this issue in sending everything to the cloud, therefore greatly increasing productivity. This may be a good idea to follow in our system.

[25] Curation and Preservation of Research Data in an iRODS Data Grid

An alternative of the precedent storage solution found (GReIC) could be iRODS. iRODS has the huge advantage to be able to be used as the backend of Fedora (one of the most widely used digital library solution). This paper exposes how to implement iRODS with an emphasis on data preservation.

REFERENCES


