Invited paper

Building a Digital Library for Epidemic Modelling

Mário J. Silva, Fabrício A. B. da Silva, Luís Filipe Lopes, Francisco M. Couto
University of Lisbon, Faculty of Sciences, LASIGE
Lisbon, Portugal

Abstract

We discuss the challenges and requirements faced in the creation of a digital library for epidemic modelling and forecasting. These are presented within the context of the Epidemic Marketplace, a distributed data management platform where epidemiological data can be stored, interlinked and made available to assist epidemiologists and public health scientists in sharing and exchanging data. We also introduce its architecture and implementation plan based on open-source tools.

Keywords

Digital libraries development, architecture, and management; DL case studies; epidemics modelling; data-intensive science

Introduction

In recent years, the availability of a huge flow of quantitative social, demographic and behavioural data spurred the interest in adopting innovative data-intensive science technologies (Hey et al. 2009). These can be used to improve disease surveillance systems with faster and more accurate outbreak detection and epidemics propagation capabilities. These capabilities depend on the availability of fine-tuned models, which require in turn accurate and comprehensive data.

We provide an overview of the architecture and design of a new digital library for scientific data analyses in epidemics modelling, called the Epidemic Marketplace (EM). In this platform, epidemiological data can be captured and curated for analysis by the scientific community. Based on open-source software, the Epidemic Marketplace is part of a computational framework for organising data for epidemic modelling and forecasting, the Epiwork project (Epiwork 2010), an e-Science initiative aimed at collecting and organizing data, creating tools and deriving knowledge to be used by epidemiologists and public health scientists.

The Epidemic Marketplace, supports the creation of large scale, data driven computational simulations of disease propagation, endowed with a high level of realism and aimed at epidemics forecasting. In addition, it offers original web data-collection schemes for integrating real-time disease incidence independently of health authorities. It is starting as an annotated catalogue of datasets of interest to epidemic modellers, which will over time into a collection of interlinked data, having some of its data elements locally stored within the realm of the Marketplace and some others available as interlinked references remotely managed.

In this paper, we start by introducing the use of metadata and ontologies for epidemic modelling and our strategies for identifyingcontrolled vocabularies and ontologies for characterising the diverse data used by epidemic modellers. We then present the software architecture of the proposed Epidemic Marketplace through the discussion of its requirements. Next, we describe the base technologies used in the construction of the infrastructure. In the conclusion, we discuss the status of the prototype implementation of the Epidemic Marketplace and our plans for fostering its use by the epidemic modellers community.

Metadata and Ontologies in Epidemic Modelling

The description of the datasets used in the models of all sorts of epidemics would requireall the necessary to propose a model capable of describing virtually every kind of information, given the diversity of factors and

the interdisciplinary of epidemiologic studies. In the study of a specific disease it is possible to have datasets describing the disease, how it spreads, clinical data about a population and so on. Data may be geo-referenced and geographic data may be necessary for the modelling of the disease transmission. Other data can be important for the study of diseases, such as genetic, socio-economic, demographic, environmental and behavioural data. The need to encompass so many areas of study will reflect on the contents of the datasets and ultimately on their metadata, calling for a data organisation supporting interlinked data (Bodenreider and Stevens 2006; Bizer in press)

Given the high diversity and heterogeneity of epidemic data involved, a common reference model based on metadata is needed. Metadata terms are being defined based on controlled vocabularies and ontology terms, and ontologies will be also used to characterize the entities and relationships among them in the managed datasets. As a result, the information model of the Epidemic Marketplace is directly defined through metadata and ontologies.Together, they will be essential in the development of epidemic modellingdigital libraries, as they make documents and other data sources accessible in a more sophisticated, structured and meaningful manner.

For example, using a specific ontology to describe a specific disease makes everybody referring to a specific disease to use the same term, making the information discovery simpler and more complete. But it also keeps the metadata text simpler, since the ontology itself contains other data that doesn't need to be inserted as metadata. For example, through an ontology of places (a geographic ontology), if we have a specific location code, we can obtain other information about that location, such as country, coordinates, altitude, city and so on.

Standards

There are several standards for the collection and management of metadata. ISO/IEC 11179 is the international standard for representing metadata for an organization in a Metadata Registry (MDR) that has been implemented by organizations in the Health domain. Several health organizations have created implementations of this MDR, such as METeOR (Australian Institute of Health and Welfare 2009). However, the DCES (Dublin Core Metadata Element Set) is the most relevant standard to our epidemic modelling e-science infrastructure, because it was conceived for describing web resources, and that is the way the Epidemic Marketplace will be primarily available(DCMI 2008). The DCES is a vocabulary of fifteen properties to be used to describe document-like files in the web. These fifteen elements are a part of a larger set of metadata vocabularies and technical specifications maintained by the DCMI (Dublin Core Metadata Initiative), the DCMI Metadata Terms (DCMI 2008b). DCMI includes, formal domains and ranges in the definitions of its properties. This means that each property may be related to one or more classes by a *has domain* relationship, indicating the class of resources that the property should be used to describe, and to one or more classes by a *has range* relationship, indicating the class of resources that should be used as values for that property (Powell et al. 2008).

The DCMI recommends the use of controlled languages whenever possible for the description of each element. However, the development of an ontology that is accepted by the whole community is a complex lengthy and costly endeavour, so it is important to reuse as much as possible existing ontology's, because they reduce costs and implementation time. In addition, adopting already used ontologies simplifies access to interlinked datasets.

The OBO (Open Biomedical Ontologies) isa repository of openly available and relevant ontologies to our problem domain (Smith 2007).We will adopt relevant ontologies from this realm and also controlled languages, such as the ones recommended by de DCMI. One example is the UMLS Metathesaurus

(Bodenrieder 2004), which is commonly used in metadata descriptions based on the Dublin Core Standards in the biomedical domain.

The DCMI suggests the use of the TGN - Thesaurus of Geographic Names (Harpring 1997) for location references. However, the use of ontologies, such as Geo-Net-PT that we developed for Portugal (Pellicer et al. 2010) or Yahoo! GeoPlanet (Yahoo, 2009)can make the annotation more exact.

We are also tracking novel services provided by INSPIRE, an European Commission initiative establishing an infrastructure for spatial information in Europe (European Commission 2007).

Strategies for Creating an Epidemic Data Metadata Model

In a first stage, the Epidemic Marketplace aims at creating a catalogue of epidemic datasets with extensive meta-data describing their main characteristics. Ontologies will play an important role in establishing the common terminology to be used in this process and to interlink heterogeneous meta-data classifications. The Epidemic Marketplace will explore a comprehensive set of relevant ontologies that besides being used to characterise datasets, will also become important datasets to epidemic modellers. Some of these are already being organised in collections. .

At a later stage, the marketplace will provide a unified and integrated approach for the management of epidemic data sources. Ontologies will have an important role in integrating these heterogeneousdata sources by providing semantic relationships among the described objects. Further on, the marketplace will include methods and services for aligning the ontologies. The aligned ontologies and annotated datasets will eventually serve as the basis for a distributed information reference for epidemic modellers, which will help further on the integration and communication among the community of epidemiologists.

To describe the epidemic datasets, it is first necessary to describe the datasets as web resources. This will be done using the DCMI terms and conventions.It will also be necessary to describe the information contained in the datasets. These descriptions constitute what health professionals and researchers will be ultimately looking for.

The level of detail of the metadata is another aspect that must be carefully designed: a low level of detail may not be able to sufficiently describe the datasets, making the right information harder to find, but a too detailed metadata scheme can turn the annotation of a specific dataset into a daunting task, hindering the acceptance of the model by the user community.In view of this, we intend to start modelling the datasets with a low level of detail, annotating the 15 standard DC elements as character data. Further down the line, we will support the extension of the DC elements annotations with semantically richer descriptions. That will be initially done with the analysis of datasets to be provided by Epiwork partners. The collaboration with these partners will enable the assessment of which level of detail will be most adequate to the epidemic modellers community.

To be useful, metadata annotation criteria have to follow a common standard, so data can be comparable and searched using similar queries. In order to obtain a standardization of the metadata annotation it is fundamental to use controlled languages as much as possible and languages for describing data structures, progressively limiting the use of free text.

To understand the metadata to be added to annotate epidemic datasets and what properties should be extended in the future for a better data representation, we have analysed a selected sample of datasets:

EM Twitter Datasets: Twitter data harvested by an initial prototype of the Data Collector module of the Epidemic Marketplace (Lopes et al. 2009). Each dataset contains tweets (messages) with disease and geographic specific keywords. It also contains, for each message, information about the author name (nickname), the source (in this case the Twitter.com service), the keywords searched, the date, the source and a possible score (assigned according to the confidence on the specific message).

US Airports Dataset: Data about the airport network of the United States. This dataset provides information about the US transportation network, containing data about the 500 US airports with most traffic. The file contains an anonymised list of connected pairs of nodes and the weight associated to the edge, expressed in terms of number of available seats on the given connection on a yearly basis.

In addition, to add more diversity to these initial datasets and start with a larger study base, we surveyed published articles in epidemiology journals for analysis and inferred the attributes of that datasets reported in those papers. Most of the studies do not provide information on how to access all the used datasets or fully describe them for the purposed of cataloguing with the detail we are envisioning. Nevertheless, this kind of survey provided insights on the metadata modelling aspects that have to be accounted for. We characterized datasets used in studies like:

Cohen et al. (2008): Analyses the relation of levels of household malaria risk with topography related humidity.

East et al. (2008): Analyses the patterns of bird migration in order to identify areas in Australia where the risk of avian influenza transmission from migrating birds is higher.

Starr et al. (2009): Introduces a model for predicting the spread of Clostridium difficile in hospital context.

Using this approach, we have annotated datasets to which we did not actually have access, but devised what would be their metadata description as DC elements, based on the information provided.

Epidemic Marketplace Architecture

The Epidemic Marketplace is composed of a set of,geographically distributed, interconnected data management nodes, sharing common data models, an authorization infrastructure and access interfaces. At each node, a set of software components implements a set of functional and non-functional requirements that characterize their performance and interfaces.

As shown in Figure , each Epidemic Marketplace node has the following four main modules:

Repository: Stores epidemic data sets and an epidemic ontology to characterise the semantic information of the data sets.

Mediator: A collection of web services that will provide access to internal data and external sources, based on a catalogue describing existing epidemic databases through their metadata using state-of-the-art semantic-web/grid technologies.

Collector: Retrieves information of real-time disease incidences from publicly available data sources, such as social networks; after retrieval, the collector groups the incidences by subject and creates data sets to store in the repository.

Forum: Allows users to organize discussions centred on the datasetsmanaged by the Epidemic Marketplace, fostering collaboration among modellers.

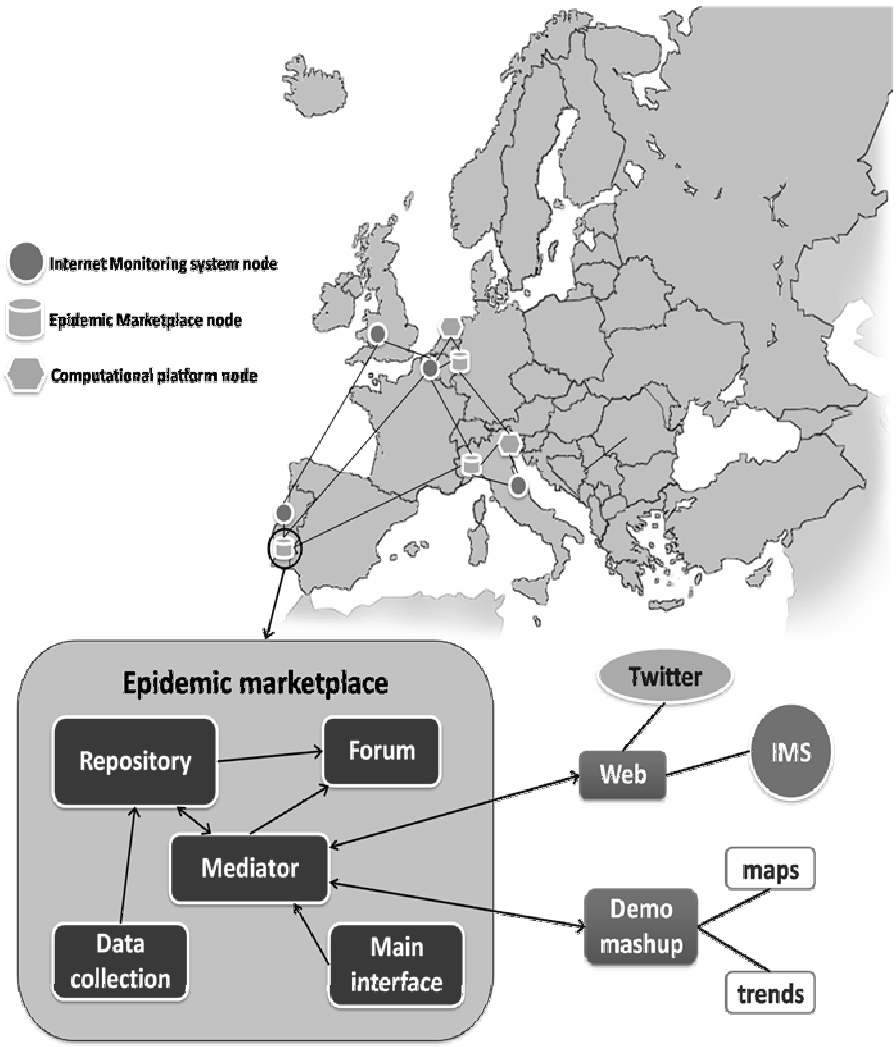

Figure 1 - An envisioned deployment of the distributed Epidemic Marketplace

The Epidemic Marketplace is a *distributed virtual repository,* a platform supporting *transparent*, seamless access to distributed, heterogeneous and redundant resources (Kuliberda et al. 2006, Ohno-Machado et al. 1997). It is a *virtual* repository because data can be stored in systems that are external to the Epidemic Marketplace, and it provides *transparent* access because several heterogeneities are hidden from its users. Data can be either stored in one or more repositories or retrieved from external data sources using authorization credentials provided by clients. Data can also be replicated among repositories to improve

access time, availability and fault tolerance. However, data replication is not mandatory; in several cases data must be stored in a single site due to, for instance, security constraints. It is worth noting, though, that any individual repository that composes the Marketplace will enable virtualized access to these data, once a user provides adequate security credentials.

In this section, we introduce the Epidemic Marketplace by presenting and discussing its main requirements.

Epidemic Marketplace Requirements

A number of projects retrieve epidemic data and make them available to users, such as Healthmap (Borownstein et al. 2006), MedISys (Mawudeku et al. 2005) and GIDEON (Gideon 2010). However, the set of requirements of the Epidemic Marketplace makes this platform quite different from previous projects. The main system requirements identified of the Epidemic Marketplace are listed below:

Support the sharing and management of epidemiological data sets: Registered users should be able to upload annotated data sets, and a data set rating assessment mechanism should be available. The annotated data set will then compose a catalogue that will be available to users.

Support the seamless integration of multiple heterogeneous data sources: Users should be able to have a unified view of related data sources. Data should be available from streaming, static and dynamic sources. All data retrieved by users or other services should be available through a common interface.

Support the creation of a virtual community for epidemic research: The platform will serve as a forum for discussion that will guide the community into uncovering the necessities of sharing data between providers and modellers. Users will become active participants, generating information and providing data for sharing and collaborating online.

Distributed Architecture: The Epidemic Marketplace should implement a geographically distributed architecture deployed in several sites. The distributed architecture should provide improved data access performance, improved availability and fault-tolerance.

Support secure access to data: Access to data should be controlled. The marketplace should provide single sign on, distributed federated authorization and multiple access policies, customizable by users.

Support data analysis and simulation in grid environments: The Epidemic Marketplace will provide data analysis and simulation services in a grid environment. Therefore, the Epidemic Marketplace should operate seamlessly with grid-specific services, such as grid security services, information services and resource allocation services.

Workflow: The platform should provide workflow support for data processing and external service interaction. This requirement is particularly important for those services that retrieve data from the Epidemic Marketplace, process it, and store the processed data back in the marketplace, such as grid-enabled data analysis and simulation services.

The main non-functional requirements that have been identified for the Epidemic Marketplace are listed below:

Interoperability: The Epidemic Marketplace must interoperate with other software. Its design must take into account that in the future, systems developed by other researchers across the world may need to query the Epidemic Marketplace catalogue for access to its datasets.

Open-source: All software packages to be used in the implementation and deployment of the Epidemic Marketplace should be open source, as well as the new modules developed specifically to the Epidemic Marketplace. An open-source based solution reduces development cost, improves software trustworthiness and reliability and simplifies support.

Standards-based: To guarantee software interoperability and the seamless integration of all geographically dispersed sites of the Epidemic Marketplace, the system will be entirely built over standards defining web services, authentication and metadata.

Repository Requirements

The objective of the Epidemic Marketplace repository is to organize the information about existing datasets. While it is expected that the datasets be deposited in the repository, it is possible to have information about specific datasets even if they are not stored at the repository. This may happen, for example, for security reasons. For these special datasets, the metadata services to be provided by the content repository will become the only alternative. The metadata repository will store information about specific datasets even if they are not in the repository. The metadata will describe the datasets in detail, including their contents, providing information about the authors, where the dataset is available and who has access to it. The main requirements of the Repository are:

Separation of data and metadata: An important architectural feature for scientific repositories in general, and also the Epidemic Marketplace, is a clear separation between data and metadata (Stolte et al. 2003). For instance, there should be a clear separation between metadata and actual data schemes, since metadata may contain information not directly available in data schemes.

Support for Meta-data standards: Extensive support for metadata standards for web resources management and processing (e.g. searching) is required. This means the adoption of Dublin Core. It is possible that only the metadata of some data sources is available through the Epidemic Marketplace, due to privacy constraints. In those cases the client should retrieve the data directly from the site hosting the data source, following directives described in the Epidemic Marketplace.

Ontology support: One step further in the deployment of the Epidemic Marketplace is to have a semantically enabled repository using ontologies for describing and structuring the data. The Epidemic Marketplace will provide a framework for the creation and development of epidemiological ontologies, openly addressing the needs of this community and fostering its active involvement(Goni et al. 1997, Fox et al. 2006).

Mediator Requirements

The Mediator is responsible for communicating with: 1) clients, which retrieve the data collections of the Epidemic Marketplace and produce dynamical trends graphs or geographical maps according to user interaction; 2) Epiwork applications, such as Internet-based Monitoring Systems (IMS) or computational platforms (CP) for simulating the propagation of diseases; 3) other data providers, such as online news wires, RSS feeds, ProMED Mail, validated official alerts (WHO) and other event generators. The main requirements of the mediator are:

Heterogeneous datasets query and search capabilities: The Mediator has to manage the access to data from many different sources, pertaining to different diseases, and in different formats, using data query or search interfaces. Besides medical information, other types of information are needed in epidemics simulations, such as geographic, sociological and about transportation networks. The data needed for an epidemic study can change significantly from disease to disease and even between studies on the same subject, depending on experimental conditions and data collection methods.

Access to "plug-in-able" resources: One important feature to be supported by the Epidemic Marketplace, in particular for external data resources, is access to "plug-in-able" resources (Kuliberda et al. 2006). These external resources provide data not stored in an internal repository and may require virtualized access. Some resources can appear and disappear unexpectedly, due, for instance, to web site unavailability. "Plug-in-able" resources enable the dynamic addition of data sources through wrappers that assure physical connection to a source and convert the gathered data to one or more of the canonical data models supported by the repository.

RESTful interface: Clients should be able to search and query datasets and corresponding metadata through RESTful interfaces.

Collector Requirements

Recent epidemiological surveillance projects are collecting data from the Internet to identify disease propagation. These systems mainly collect data from pre-selected data sources somehow related to the subject. However, other sources, like social networks and search engine query data, may present early evidence of an infection event and propagation (Ginsberg et al. 2008). Given the increasing popularity of social networks, we can find a large amount of personal information in real time, which canhelp in detect earlier the beginning or the propagation of an epidemic event. The main requirements of the Collector are:

Active data harvesting: The Collector should actively harvest data about putative infections by automatically retrieving infection alerts from the Webusing a focused web crawler (Chakrabarty, 1999), subscription of newsfeeds and email services.

Passive data collection: The data collector should also be able to receive information directly from online users accessing the Epidemic Marketplace using data upload forms or deposited from Internet Monitoring Systems (van Noort et al. 2007).

Local storage capability: All collected data should be physically stored in at least one site of the Epidemic Marketplace. This is important since the data may be nolonger available from its source after some time. Data should be organized as datasets following partitioning criteria meaningful to epidemic modellers.

Forum Requirements

The Epidemic Marketplace will serve as an exchange platform for connecting modellers, who search for input data for calibrating and evaluating and their models, and providers, who seek the help of modellers for obtaining analyses and interpret their data. Therefore, its user community requires an online meeting point for discussions about the data collections and for uncovering the data sharing requirements among providers and modellers. This will promote collaborations, through direct trustful sharing of data within the communities and establishment of consensus agreements between modellers and data providers on sharing data for epidemics modelling. The results will be reported to EU-agencies, such as the ECDC and the EMCDDA, as a contribution to setting European standards for sharing epidemic data. The main requirements of the Epidemic Marketplace Forum are:

Group-oriented discussions with access restrictions: Every discussion should be associated with a group of users. A user uploading a new dataset into the EM Repository defines membership and access restrictionsto the corresponding online discussion group.

Support for distributed authentication: As it is the case with the Mediator, clients must authenticate to at least one site of the Epidemic marketplace to Access the forum. The same set of credentials for a given client should be accepted by any instance of the Epidemic Marketplace. After authentication, the client is redirected to the Epidemic Marketplace site hosting the discussion, if the user is included in the associated group access list.

Epidemic Marketplace Implementation

A first prototype version of the Epidemic Marketplace is under construction. The initial version is already in use by the project members and implements several of the main features of the outlined architecture, including data management and sharing support and secure access to data, and is currently being populated with epidemic datasets and their metadata descriptions.

Several open-source tools and open standards are being used in the Epidemic Marketplace implementation and deployment process. We selected Fedora Commons (Lagoze et al. 2006) for the implementation of the main features of the repository. Access control in the platform implements the XACML (OASIS 2010), LDAP (Tuttle et al. 2004) and Shibolleth(Internet2 2010) standards. We started with a front-end based in Muradora(Nguyen and Dalziel 2008), which is now being replaced by a new front-end based in the Drupal content management system (Byron et al. 2009).

The prototype is intended as an initial "proof-of-concept", but it already shows the limitations spanning from annotating the datasets using free text in the description fields. This might be useful for some search applications, but quickly becomes very subjective. For this kind of informal annotation, we will provide a much simpler model, inspired on web2.0 "tags." EM users will be able to freely annotate their datasets using their own terminologies (also dubbed as "folksonomies").

We are now working on the extension of this model, using the recent DCMI terms and other specific extensions. To do so, we are completing the Epidemic Marketplace Dublin Core application profile, where the specific extensions needed and the identification of controlled languages that can be used in order to avoid subjectivity and implement a high standardization level are being identified. We are also relying on the community to guide its iterative development.

Conclusion

The development and implementation of the Catalogue of the Epidemic Marketplace is tightly connected with the population of the Repository with different kinds of epidemic datasets and discussions for better understanding how a metadata description can be made as exact and complete as needed, and still be useful and acceptable to the occasional visitor who deposits a dataset or wants to annotate it.

We have started using existing ontologies, such as the UMLS. Our goal is to contribute to making ontologies widely accepted by the Epidemiological community and ensuring their sustainable evolution, by replicating the success of similar initiatives, such as the Gene Ontology in Molecular Biology (Ashburner et al. 2000).

The method of scanning published epidemic modelling studies, extracting references (explicit and implicit) to the described datasets, and then inferring the meta-data descriptions they should have, is being very useful, since it makes it possible to understand the variety of data used in epidemic studies, how it is related, and understand the difficulties that this community would experience in providing it. Moreover, these inferred annotations can also be used as examples to new Epidemic Marketplace users with no previous metadata definition experience. We believe that thismay spawn the development of increasingly richerand accurate metadata characterisations of epidemic datasets.

The biggest challenge that lies ahead is how to motivate the community to populate the Epidemic Marketplace. We will soon be facing an instance of the classic "chicken-and-egg problem,"where the prototype is not perceived as an attractive resource because it has not a rich collection of datasets, and it hasn't more datasets because the community does not perceive its potential. Our Epiwork partners who have been active in creating models using real world data that they have collected over the years will have a key role.

Another strategy involves the active collection of data and updates to datasets from the web, for automatic annotation or archival into the Epidemic Marketplace. Ourcurrent prototype has been collecting data from Twitter on a daily basis. It is worth noting that the EM not only collects the data, but also stores them. This is important because messages in Twitter are only available for one month. As we are periodically assembling these messages into semantically annotated data collections in the Repository, they could become a useful resource for researchers modelling the spreading of diseases. In the future we could correlate the predictions made from the data in these collections with official statistics and assess its accuracy. Previous work with web search logs data, which are private, has shown how effective these short texts can be for predicting epidemic outbreaks when the date and location of their authors can be traced (Ginsberg et al. 2008).

We will welcome any other organizations willing to participate in the Epidemic Marketplace after the stress tests that are underway and the software reaches beta-level quality. We will also make the full source code available as Open Source and encourage the development of extensions. Later on, we expect to publish the first integrated models providing integrated views of both internally and externally stored data together with a catalogue of available epidemiological data.

Acknowledgements

The Epiwork project is a large multi-organizational initiative and the ideas that lead to the Epidemic Marketplace are the result of discussions with its developers and participants, in particular our students Hugo Ferreira, JoãoZamite e Patrícia Sousa.Epiwork is funded by the European Commission under the Seventh Framework Programme (Grant # 231807). We also thank FCT (Portuguese research funding agency) for its LASIGE Multi-annual support.

References

Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin GM, Sherlock G. 2000.
Gene o The objective of the Epidemic Marketplace repository is to The objective of the Epidemic Marketplace repository is to **ntology: tool for the unification of biology.**
Nature Genetics 25(1):25-29.

Australian Institute of Health and Welfare.2010
METeOR, Metadata Online Registry.
http://meteor.aihw.gov.au/content/index.phtml/itemId/181162 (accessed September 2009).

Bizer C, Heath T and Berners-Lee T.in press.
Linked Data - The Story So Far.
International Journal on Semantic Web and Information Systems, Special Issue on Linked Data.

Bodenreider, O. 2004.
The Unified Medical Language System (UMLS): integrating biomedical terminology.
Nucleic Acids Research 32(Database issue):D267-270.

Bodenreider, O and Stevens, R. 2006.
Bio-ontologies: current trends and future directions.
Briefings in Bioinformatics 7(3):256-274.

Brownstein, JS, Freifeld, CC. 2008.
HealthMap: the development of automated real-time internet surveillance for epidemic intelligence.
Euro Surveillance 12: E071129 071125.
http://www.eurosurveillance.org/ew/2007/071129.asp#5. (accessed January 2010).

Byron A, Berry A, Eaton J, Haug N, Walker J and Robbins J. 2008.
Using Drupal.
O'Reilly Media, Inc.560 pages. ISBN-10: 0596515804l.

Chakrabarti S, van den Berg M and Dom B. 1999.
Focused crawling: a new approach to topic-specific Web resource discovery.
Proceedings of the 8th International World Wide Web Conference, pp. 545-562.

Cohen, J.M, Ernst, K.C., Lindblade K.A., Vulule J.M., John C.C. and Wilson M.L. 2008.
Topography-derived wetness indices are associated with household-level malaria risk in two communities in the western Kenyan highlands.
Malaria Journal, 7: 40.

Dublin Core Metadata Initiative. 2008.
Dublin Core Metadata Element Set, Version 1.1 (ISO standard 15836-2003).
http://dublincore.org/documents/dces/ (accessed on January, 2010)

Dublin Core Metadata Initiative. 2008b.
DCMI Metadata Terms.
http://dublincore.org/documents/dcmi-terms/ (accessed on January, 2010)

East I.J., Hamilton S. and Garner M.G. 2008.
Identifying areas of Australia at risk of H5N1 avian influenza infection from exposure to migratory birds: a spatial analysis.
Geospatial Health 2(2):203-213.

Epiwork Project. 2010.
Developing the framework for an epidemic forecast infrastructure.
http://www.epiwork.eu (accessed January 2010).

European Commision. 2007.
INSPIRE Directive.
http://inspire.jrc.ec.europa.eu/ (accessed January 2010)

European Comission. 2010
MedISys.
http://medusa.jrc.it/medisys/aboutMediSys.html (accessed January 2010).

Fox P, McGuinness D, Middleton D, Cinquini L, Anthony Darnell J, Garcia J, West P, Benedict J, Solomon, S. 2006.
Semantically-Enabled Large-Scale Science Data Repositories.
Proceedings of the 2006 International Semantic Web Conference (ISWC), LNCS vol. 4273.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2008.
Detecting influenza epidemics using search engine query data.
*Nature*457, 1012-1014 (19 February 2009)

Goni A, Mena E, Illarramendi A. 1997.
Querying Heterogeneous and Distributed Data Repositories using Ontologies.
Proceedings of the 7th European-Japanese Conference on Information Modelling and Knowledge Bases (IMKB'97).

Harpring, P. 1997.
Proper words in proper places: The thesaurus of geographic names.
MDA Information, 2, pp. 5-12.

Hey T, Tansley S, Tolle K (editors). 2009.
The Fourth Paradigm: Data-Intensive scientific Discovery.
Microsoft Research.

GIDEON Informatics. 2010.
GIDEON - the Global Infectious Disease & Epidemiology Network
http://www.gideononline.com/ (accessed January 2010).

Internet2. 2010
Shibboleth.
http://shibboleth.internet2.edu (accessed January 2010).

Kuliberda K, Blaszczyk P, Balcerzak G, Kaczmarski K, Adamus R, Subieta K. (2006).
Virtual Repository Supporting Integration of Pluginable Resources.
Proceedings of the IEEE 17th International Conference on Databases and Expert Systems Applications (DEXA'06).

Lagoze C, Payette S, Shin E, Wilper C. 2006.
Fedora: an Architecture for Complex Objects and their Relationships.
*International Journal on Digital Libraries.*Vol. 6(2), pp 124-138.

Lopes LF, Zamite JM, Tavares BC, Couto FM, Silva F and Silva MJ. 2009.
Automated Social Network Epidemic Data Collector.
In Luis Rodrigues and Rui Lopes (Eds.) *Actas do INForum - Simpósio de Informática 2009*. Faculdade de CiênciasdaUniversidade de Lisboa.Setembro de 2009. ISBN: 978-972-9348-18-1.

Mawudeku A and Blench M.
Global Public Health Intelligence Network (GPHIN).
In *Proceedings of the 7th Conf. of the Association for Machine Translation in the Americas.*
http://www.mt-archive.info/MTS-2005-Mawudeku.pdf (accessed February 2008).

Nguyen C and Dalziel J. 2008.
Muradora: A Turnkey Fedora GUI Supporting Heterogeneous Metadata, Federated Identity, and Flexible Access Control.
in*Proceedings of the Third International Conference on Open Repositories*, 2008.

OASIS. 2010.
OASIS eXtensible Access Control Markup Language (XACML).
http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml (accessed January 2010).

Ohno-Machado L, Boxwala A, Ehresman J, Smith D, Greenes R. 1997.
A Virtual Repository Approach to Clinical and Utilization Studies: Application in

Mammography as Alternative to a National Database.
In *Proceedings of the 1997 AMIA Annual Symposium*.

Pellicer FJ, Chaves M, Rodrigues C, Silva MJ. 2009.
Geographic ontologies production in Grease-II.
Technical Report TR 09-18. University of Lisbon, Faculty of Sciences, LASIGE.doi:10455/3256.

Powell A, Johnston P., Baker T. 2008.
Domains and Ranges for DCMI Properties.
http://dublincore.org/documents/2008/01/14/domain-range/ (accessed January 2010).

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W et al. 2007.
The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.
Nature Biotechnology 25 (11): 1251-5. doi:10.1038/nbt1346. PMID 17989687

Starr J.M., Campbell A., Renshaw E., Poxton I.R., and Gibson G.J. (2009).
Spatio-temporal stochastic modelling of Clostridium difficile.
The Journal of Hospital Infection 71(1):49-56.

Stolte E, von Praun C, Alonso G, Gross T. 2003.
Scientific Data Repositories – Designing for a Moving Target.
in*Proceedings of the 2003 ACM SIGMOD international Conference on Management of Data,* pp 349-360.

Tuttle S, Ehlenberger A, Gorthi R, Leiserson J, Owen N, Ranahandola S, Storrs M and Yang C. 2004.
Understanding LDAP design and implementation.
IBM International Technical Support Organization, 2nd ed.

vanNoort SP, Muehlen M, Rebelo de Andrade H, Koppeschaar C, Lima Lourenço JM, Gomes MG. 2007.
Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe.
Euro Surveillance. 2007;12(7):pii=722.
http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=722

Yahoo, Inc. 2009.
Yahoo! GeoPlanet.
http://developer.yahoo.com/geo/geoplanet/ (accessed on January 2010).