

UNIVERSIDADE DE LISBOA

Faculdade de Ciências

Departamento de Informática



A METADATA MODEL FOR THE
ANNOTATION OF EPIDEMIOLOGICAL DATA

Luis Filipe Vieira da Silva Lopes

Mestrado em Tecnologias da Informação Aplicadas às
Ciências Biológicas e Médicas

2010

UNIVERSIDADE DE LISBOA

Faculdade de Ciências

Departamento de Informática



A METADATA MODEL FOR THE
ANNOTATION OF EPIDEMIOLOGICAL DATA

Luis Filipe Vieira da Silva Lopes

Trabalho de Projecto orientado pelo Prof. Doutor Fabrício Alves Barbosa da Silva e
co-orientado pelo Prof. Doutor Mário Jorge Costa Gaspar da Silva

Mestrado em Tecnologias da Informação Aplicadas às
Ciências Biológicas e Médicas

2010

Resumo

Esta dissertação apresenta um modelo de metadados para integração, gestão e partilha de dados epidemiológicos. O modelo incorpora elementos do Dublin Core, um standard para anotação de metadados largamente usado na internet. São também incluídos outros elementos de forma a melhor estruturar os termos Dublin Core, além de novos elementos para a descrição de conceitos epidemiológicos, ou relacionados, de uma forma mais específica.

O modelo foi desenvolvido para a *Epidemic Marketplace*, uma plataforma de gestão e integração de dados para sistemas de modelação epidemiológica em desenvolvimento no âmbito do projecto de investigação EPIWORK.

O repositório digital da *Epidemic Marketplace* foi construído fazendo uso do modelo de metadados desenvolvido neste trabalho, sobre a plataforma *Fedora Commons*, usando o software *Muradora* como interface. A anotação de recursos é assistida através do uso de listas baseadas em vocabulários controlados, menus de ajuda e preenchimento automático de metadados. O uso de vocabulários controlados, gerados frequentemente a partir de bases de termos ontológicos, é essencial para melhorar a qualidade da representação semântica dos metadados e facilita a sua interpretação automática.

Palavras-chave: Biblioteca Digital, Modelo de metadados, EPIWORK, Epidemic Marketplace

Abstract

This thesis presents a metadata model for integration, management and sharing of epidemiological data. The model incorporates elements from the Dublin Core metadata standard along with new metadata elements to extend and structure the Dublin Core terms. It also includes new elements for the description of the specificities of the epidemiological information.

The model was developed for the *Epidemic Marketplace*, a digital library for epidemic modeling systems under development within the EPIWORK research project. The deployed digital repository of the Epidemic Marketplace was implemented based on this model, using the *Fedora Commons* platform with *Muradadora* as front-end. The annotation of resources is assisted by controlled vocabularies, help menus and automatic filling of metadata. The use of controlled vocabularies, often created from ontologic term lists, keeps metadata consistent, improves its semantics, and facilitates the automatic interpretation of metadata.

Keywords: Digital Library, Metadata Model, EPIWORK, Epidemic Marketplace

Resumo Alargado

Estudos epidemiológicos são fundamentais para a identificação de factores que afectam o aparecimento e transmissão de doenças. Por outro lado técnicas de modelação e simulação epidemiológica são necessários para prever o comportamento de doenças e eventos epidemiológicos. Para que a modelação epidemiológica possa ter um elevado poder de previsão necessita ter acesso a grandes quantidades de dados. Estes precisam ainda de ser detalhados e cobrir os mais variados aspectos da doença e factores que a afectam.

Dados epidemiológicos, acerca de doenças que afectam o Homem são constantemente obtidos em estudos científicos e na prática clínica, no entanto, esses dados são de difícil acesso devido sobretudo a questões legais relacionadas com privacidade. Enquanto não forem aceites e postos em prática mecanismos que permitam uma utilização mais intensiva desses dados, tais como processos de anonimização e pseudoanonimização, esses dados continuarão na sua maioria inacessíveis.

Nos últimos anos a internet tem sido explorada como alternativas para a obtenção de dados epidemiológicos, a partir de dados voluntariamente colocados online por utilizadores. Algumas das fontes utilizadas têm sido, por exemplo, termos inseridos em motores de pesquisa (*Google Flu Trends*), redes sociais (ex.: *Twitter*), sites de notícias (*Healthmap*) ou mesmo dados fornecidos voluntariamente por utilizadores registados (*Internet Monitoring Systems*).

O *Epidemic Marketplace* é uma plataforma de gestão de dados epidemiológicos integrada no projecto Europeu *Epiwork*. Este projecto visa criar um sistema integrado para a previsão de eventos epidemiológicos. Neste sistema, o *Epidemic Marketplace* estará responsável pelo armazenamento, gestão e integração de dados. O *Epidemic Marketplace* deverá para tal fornecer ferramentas para que utilizadores e aplicações, tais como sistemas de monitorização epidemiológica baseadas na internet ou sistemas computacionais para análise de dados e modelação epidemiológica, possam aí depositar e aceder a dados.

O *Epidemic Marketplace* é constituído por quatro módulos: 1) um repositório digital, para gestão de recursos digitais; 2) o *Mediator*, que providencia *webservices* para o acesso automático ao repositório, permitindo a manipulação dos recursos por aplicações fornecendo serviços para upload, busca e consulta de recursos ou dos seus metadados; 3) o *MEDCollector*, um aplicação capaz de recolher dados de forma activa e passiva de

diversas fontes na internet; 3) o fórum, onde os utilizadores do *Epidemic Marketplace* podem discutir e formar uma comunidade online, de forma a discutir o seu trabalho, recursos aí armazenados, ou mesmo criar novas colaborações.

O objectivo deste trabalho foi o desenvolver um modelo de metadados para o *Epidemic Marketplace*. Este modelo visa reforçar a integração, gestão e partilha de recursos epidemiológicos e foi concebido tendo em conta as necessidades específicas para a anotação deste tipo de recursos.

Durante este projecto, um repositório digital foi implementado usando o software *Fedora Commons* e *Muradora*, que correm sobre o servidor web *Apache Tomcat*. O *Fedora Commons* funciona como *backend* fazendo a gestão dos recursos submetidos enquanto que o *Muradora* disponibiliza um interface web para o acesso de utilizadores aos recursos armazenados. Outras aplicações são usadas para complementar o sistema fornecendo algumas das funcionalidades necessárias: i) o *OpenLDAP*, para armazenar os dados de registo de utilizadores, tais como a credenciais para aceder ao sistema; ii) o *Solr*, um servidor de busca que corre sobre o *Tomcat*; iii) o *Melcoe-PDP*, para aplicação de políticas de acesso baseadas em XACML, e que faz uso da base de dados *Oracle DBXML* para o seu armazenamento.

O repositório digital do *Epidemic Marketplace* oferece uma interface web que, mediante registo, permite aos utilizadores consultar os recursos aí armazenados e submeter os seus próprios recursos. Os recursos podem armazenados fisicamente ou simplesmente referenciados e descritos com metadados. Alguns tipos de recursos que podem ser armazenados, ou referenciados, são: datasets, documentos, recursos web ou eventos.

O acesso aos recursos é providenciado a utilizadores registados, de acordo com políticas de acesso definidos em ficheiros XACML. O sistema permite definir o acesso a recursos, quer a nível de colecção quer a nível de recursos individuais, através da autorização a grupos utilizadores.

Os recursos submetidos deverão ser anotados com metadados, de forma a possibilitar uma melhor organização e de forma a melhorar a capacidade de busca no repositório. O modelo de metadados desenvolvido neste trabalho foi baseado e usa elementos do Dublin Core, um standard para anotação de metadados largamente usado na internet. São também incluídos novos elementos de forma a expandir e a melhor estruturar o esquema de metadados. Esse esquema contém elementos específicos para a descrição de recursos epidemiológicos, geográficos, demográficos, socioeconómicos e ambientais. Os recursos epidemiológicos são descritos no set de sub-elementos de metadados do elemento

epidemiological. Este sub-set de elementos permite a descrição de conceitos epidemiológicos que poderão ser cobertos pelos recursos armazenados, como por exemplo a(s) doença(s) em questão, o hospedeiro da doença, o vector e o agente patogénico

O modelo de metadados identifica ainda diversas fontes de informação, tais como taxonomias, thesaurus ou ontologias, para a padronização e enriquecimento semântico de metadados. Essas fontes de informação poderão ser usadas para a criação de vocabulários controladas que irão contribuir para manter os metadados mais consistentes.

Para o desenvolvimento deste modelo de metadados foi necessário testá-lo continuamente, através da anotação de recursos contidos no repositório. Devido ao número de recursos armazenados no repositório ser ainda baixo, recorreu-se também à análise de artigos científicos de epidemiologia, de forma a identificar possíveis datasets usados nesses estudos. Assim, para o testar o modelo em desenvolvimento, usou-se um set composto por datasets armazenados no repositório e por datasets identificados em artigos científicos de epidemiologia. A abertura do *Epidemic Marketplace* ao público, no final de Setembro de 2010, irá potenciar um aumento do número de recursos armazenados e uma melhor avaliação e refinamento do modelo.

Apesar da utilidade dos metadados para a gestão de recursos, facilitando a sua organização e busca, um dos grandes problemas associados ao seu uso, é a resistência dos utilizadores na criação desses metadados ou a inserção de metadados inadequados. Alguns mecanismos foram implementados de forma a evitar estes problemas, como por exemplo: listas controladas, menus de ajuda e preenchimento automático de metadados.

O uso de listas controladas de termos, baseadas em vocabulários controlados, é útil pois evita a necessidade de escrever e facilita a compreensão por parte do utilizador do tipo de informação requerida. Além disso, faz com que todos os utilizadores usem o mesmo termo para descrever um conceito e evita outros problemas tais como erros ortográficos, o que promove a consistência dos metadados.

Menus de ajuda foram também implementados e são um apoio ao utilizador que desta forma poderá ter uma explicação mais detalhada caso tenha dúvidas sobre que tipo de e informação é esperada em determinado campo de metadados.

Por fim, o preenchimento automático de metadados foi implementado para os campos, de metadados "title", "author", "organisation" and "publisher". O campo "title" é preenchido automaticamente a partir do nome do ficheiro submetido. Por outro lado, dados referentes à identificação do autor, organização e editor são preenchidos automaticamente

com os dados do utilizador que submete o recurso, usando informação mantida na base de dados LDAP.

Assim, neste trabalho foi implementado um repositório digital, que permite o armazenamento e gestão de recursos epidemiológicos. Foi também desenhado esquema de metadados que permite uma anotação específica e estruturada de recursos epidemiológicos ou relacionados. O modelo de metadados foi implementado no repositório digital, através de um formulário para anotação de recursos com metadados, seguindo as especificações do modelo desenhado. Por fim, foram desenvolvidos vários mecanismos para tornar o processo de anotação mais fácil e rápido, incluindo a criação de menus de ajuda, o uso de listas controladas de termos e mecanismos automáticos de criação de metadados.

Acknowledgments

I would like to thank my advisor, Dr. Fabrício Silva, for his support and close guidance throughout this project, especially needed due to my inexperience in the informatics area.

I would also like to thank Dr. Mário Silva, co-advisor of this project, for his guidance and helpful feed-back.

I would also like to thank all the other LASIGE members with whom I have closely worked, in the Epiwork project for over one and a half years: Dr. Francisco Couto, João Zamite, Patrícia Sousa and Hugo Ferreira.

I want to thank my family, especially Paula and David for their love and for always being there and for putting up with me.

I'm also grateful to my XLDB colleagues, who have always been available to help, for their company and for providing a very nice work environment. To Cátia Machado, for her help, especially in the integration with the XLDB group and for the revision of this thesis. To Bruno Tavares, for all the support throughout the Master's (all those programming issues) and for being a good friend. To Hugo Bastos, for always being available to help and for his friendship. To Francisco Lopez-Pellicer, for his help and his sympathy during his stay at LASIGE. And also all the other group members who have contributed to make my stay at LASIGE a pleasant one: Ana Teixeira, Carlos Costa, Cátia Pesquita, Daniel Faria and Tiago Grego.

Contents

Resumo	3
Abstract	5
Resumo Alargado	7
Contents	13
Index of Figures	15
Index of Tables	17
Acronyms	19
Chapter 1 Introduction	21
1.1 Epidemic Marketplace	22
1.2 Motivation	25
1.3 Objective	26
1.4 Methodology	26
1.5 Results	27
1.6 Organization of the document	27
Chapter 2 Related Work on Data Management on the Internet	29
2.1 Metadata models	30
2.2 Metadata standards	31
2.3 Metadata registries	32
2.4 Controlled vocabularies and Encoding schemes	33
2.5 Open Archives Initiative	35
2.6 Digital repositories	35
Chapter 3 Epidemic Marketplace Digital Repository	37
3.1 Setup of the digital repository	38
3.2 Repository deployment	39
3.3 Hardware setting	40
Chapter 4 Metadata Model for the Epidemic Marketplace	43
4.1 Requirements	44
4.2 Application profile	45
4.3 Definition of metadata elements	45
4.4 Application profile organization	46

4.5	Metadata schema description.....	48
4.6	Resource type and metadata elements	51
Chapter 5 Implementation and Evaluation of the EM Metadata Model.....		57
5.1	Form design and implementation	58
5.2	Metadata creation support	63
5.3	Testing and validating	64
5.4	Analysis of datasets for the design of the metadata model.....	65
5.4.1	Twitter datasets	65
5.4.2	US Airports Dataset	69
5.4.3	Article by Cohen and coworkers	71
5.4.4	Article by East and coworkers	74
5.4.5	Article by Eubank and coworkers.....	77
Chapter 6 Conclusions		81
References.....		83
Appendix.....		89
A.1- EM Property elements based on DCTERMS.....		89
A.2- EM metadata elements defined locally		98

Index of Figures

Figure 1- The main components of the Epiwork project and how they integrate to achieve a common objective.....	22
Figure 2- Schematic representation of the distributed Epidemic Marketplace and its integration with external applications.....	23
Figure 3- The digital repository and its integration with the other EM modules and external applications.	24
Figure 4- The EM repository main page.	40
Figure 5- XML of the EM metadata schema.	49
Figure 6- Types of resources managed by the EM.	52
Figure 7- Metadata elements to be used to describe resources of the <i>dataset</i> type. ..	53
Figure 8- Metadata elements to be used to describe resources of the <i>document</i> type.	53
Figure 9- Metadata elements to be used to describe a resource of the type “web”	54
Figure 10- Metadata elements to be used to describe a resource of the type <i>event</i> . ..	55
Figure 11- Metadata elements to be used to describe a resource of the type <i>software</i>	55
Figure 12- Metadata elements required for a general description of the resource.....	58
Figure 13- Metadata elements required for the identification of authors and publisher of the resource.	59
Figure 14- Metadata elements to describe the coverage.	60
Figure 15- Metadata block to define the source of the resource.	60
Figure 16- Metadata elements for an advanced description of resource contents.	61
Figure 17- Group of metadata elements to describe a bibliographic reference, for the resource.....	62
Figure 18- Metadata elements to identify the owner and include the copyright or disclaimer of the resource.....	62
Figure 19- Annotation of a Data Collector dataset, using the DC metadata schema. 66	
Figure 20- Annotation of a Data Collector dataset, using the EM metadata schema. 67	
Figure 21- Annotation of a US airport dataset, using the DC metadata schema.	69
Figure 22- Annotation of a US airport dataset, using the EM metadata schema.	70

Figure 23- Metadata annotation of a dataset, identified in the paper by Cohen and coworkers, using the DC metadata schema.....	72
Figure 24- Metadata annotation of a dataset, identified in the paper by Cohen and coworkers, using the EM metadata schema.	73
Figure 25- Metadata annotation of a dataset, identified in the paper by East and coworkers, using the DC metadata schema.....	75
Figure 26- Metadata annotation of a dataset, identified in the paper by East and coworkers, using the EM metadata schema.	76
Figure 27- Metadata annotation of a dataset, identified in the paper by Eubank and coworkers, using the DC metadata schema.....	78
Figure 28- Metadata annotation of a dataset, identified in the paper by Eubank and coworkers, using the EM metadata schema	79

Index of Tables

Table 1- Table defining the Vocabulary Title, the Namespace Name and the Prefix used to abbreviate the Namespace name.	46
Table 2- The description of each metadata element used in this schema will be done according to the guidelines in this table (adapted from [80])......	47
Table 3- Index of metadata elements.	47
Table 4- In the DC schema the disease covered by the dataset is identified in free text using non-specific metadata fields, while in the EM schema it is identified in a specific metadata field making use of a controlled vocabulary.	68
Table 5- Spatial and temporal coverage annotated using the DC and the EM metadata schemas.....	68
Table 6- How the information about people mobility can be described in the DC and the EM schemas.....	71
Table 7- Description of demographic, environmental, epidemiological, geographical and socio-economical data in the DC and EM schemas.....	74
Table 8- Description of demographic, epidemiological and geographical data in the DC and EM schemas.	76
Table 9- Description of epidemiological, geographical and socio-economical data using the DC and EM schemas.....	77

Acronyms

ACGT	Advancing Clinico-Genomic Trials on Cancer
API	Application Programming Interface
BCP47	Best Current Practice 47
caBIG	Cancer Biomedical Informatics Grid
CDC	Centers for Disease Control and Prevention
CHDR	Clinical Data Repository/Health Data Repository
DC	Dublin Core
DCAP	Dublin Core Application Profile
DCMES	Dublin Core Metadata Element Set
DCMI	Dublin Core Metadata Initiative
DCTERMS	Dublin Core Metadata Initiative Metadata Terms
DOI	Digital Object Identifier
EM	Epidemic Marketplace
HTML	HyperText Markup Language
ICD	International Classification of Diseases
IMS	Internet Monitoring System
ISBN	International Standard Book Number
ISSN	International Standard Serial Number
JSP	Java Server Pages
LDAP	Lightweight Directory Access Protocol
MARC	MAchine-Readable Cataloging
MIME	Multipurpose Internet Mail Extensions
MIMEtype	MIME Media Type
MODS	Metadata Object Description Schema
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAI-ORE	Open Archives Initiative Object Reuse and Exchange
PHDR	Propel Population Health Data Repository
PubmedID	Pubmed IDentification number
RDF	Resource Description Framework
UMLS	Unified Medical Language System

URI	Universal Resource Identifier
URL	Universal Resource Locator
XACML	eXtensible Access Control Markup Language
XML	eXtensible Markup Language
XSL	Extensible Stylesheet Language

Chapter 1

Introduction

Epidemiological studies are fundamental to identify factors affecting disease onset and transmission. Epidemiological modeling is necessary to forecast disease behavior and transmission in order to predict epidemics and the effect of disease control measures. These types of studies need large amounts of detailed data about diseases and the factors affecting their onset and/or transmission.

The early detection of infectious disease outbreaks is fundamental for the efficient intervention of public health authorities and for the application of disease control measures [1]. Therefore, there is a constant quest for new surveillance methods, capable of decreasing the gap between disease outbreak and its detection [2].

In the last years some of these epidemiologic monitoring systems have been developed based on internet technologies. One of these systems, Gripenet, is an internet monitoring system (IMS) based on voluntary user reports [3], which is part of a network of flu monitoring systems first launched in the Netherlands [4,5].

Google Flu Trends, developed by Google, is another internet based system for epidemiological surveillance. Google's system uses search engine query data to preview influenza trends. It has been shown to relate tightly with official statistics, in the US, with the advantage of detecting disease trends earlier than official statistics [6].

Healthmap collects, and displays in a world map, information about diseases gathered from several sources, such as official alerts, the ProMED-mail newsletter and news sites [7].

The development of these approaches based on internet technologies has contributed to the continuous increase of epidemiological data available. However, those data are neither centralized nor organized in order to be easily found and shared among scientists and health professionals.

1.1 Epidemic Marketplace

Epiwork is a European project that comprises researchers from twelve different organizations, from eight different countries [8]. Among these is the LASIGE (Large Scale Informatics Systems Laboratory), from Lisbon's University Faculty of Sciences, where this master's project took place. One of the tasks in which LASIGE researchers are involved is the development of the Epidemic Marketplace (EM), available at <http://epiwork.di.fc.ul.pt> [9].

The EM is Epiwork's information platform for the integration, management and sharing of epidemiological data. This platform will interoperate with a computational modeling platform and will store data derived from internet monitoring systems. It will also provide a venue for the discussion of epidemiologic modeling issues. Together these platforms aim to provide tools for data management, epidemiological modeling and forecasting (Figure 1).

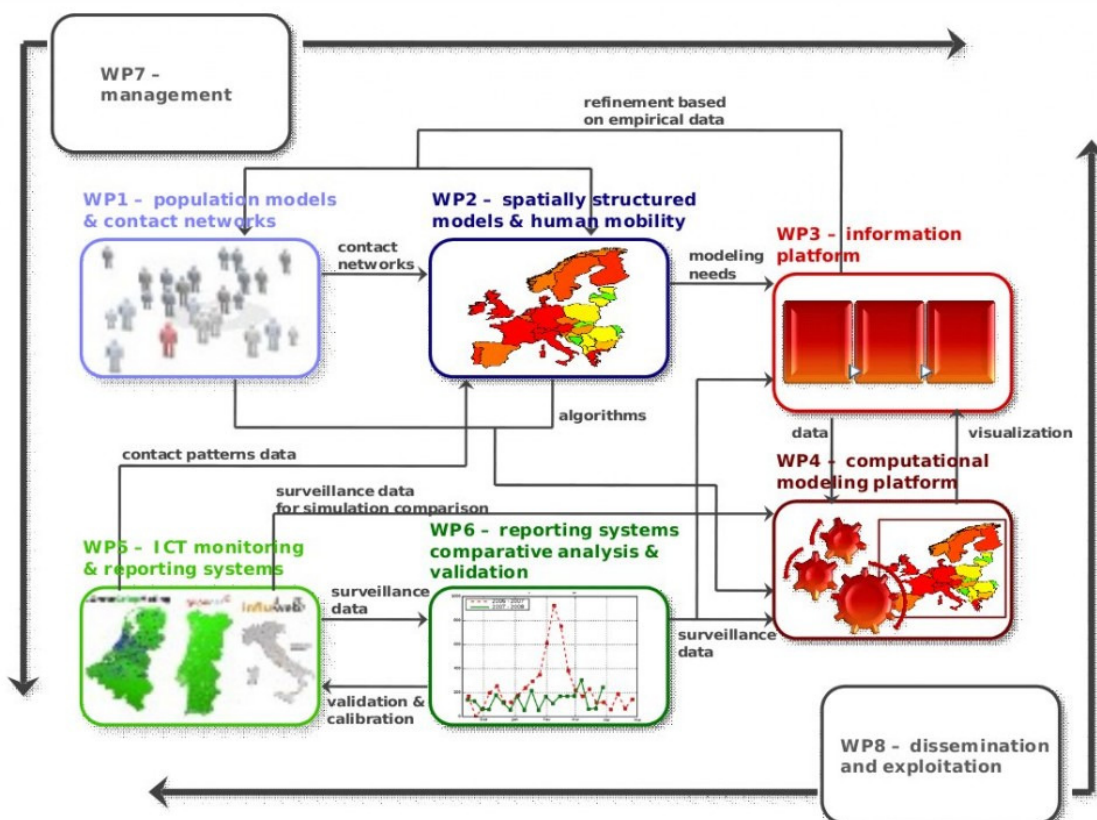


Figure 1- The main components of the Epiwork project and how they integrate to achieve a common objective. The Epidemic Marketplace is represented in WP3, providing an information platform that interacts with a computation modeling platform (WP4), stores surveillance data processed in WP6, and provides a discussion venue epidemiological modeling issues. Image from www.epiwork.eu/the-project/.

The EM is envisioned as a distributed platform, where several nodes can be implemented at different locations, forming a network and interacting with local internet monitoring systems or computational platforms (Figure 2).

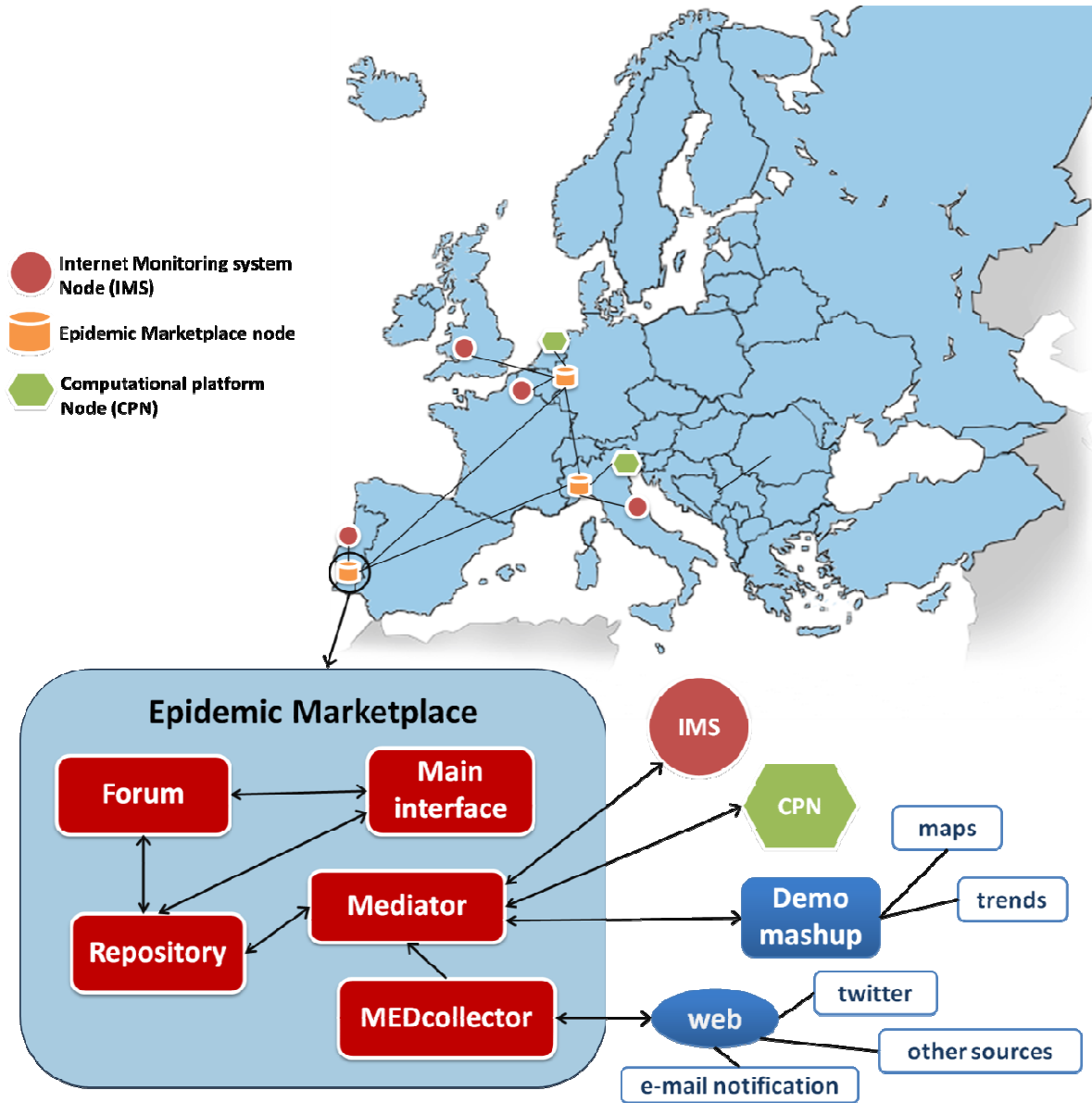


Figure 2- Schematic representation of the distributed Epidemic Marketplace and its integration with external applications.

The EM is composed by several modules:

- 1- A digital repository of datasets and other resources of interest primarily to researchers and health professionals in the field of epidemiology. The repository can be accessed at <http://epiwork.di.fc.ul.pt/muradora>.

- 2- The MEDCollector [10], which replaces the Data Collector [11], collects epidemiological data from internet sources, through the definition of workflows. Among other sources, it collects messages containing disease keywords from Twitter. The MEDCollector can be accessed at <https://epiwork.di.fc.ul.pt/medcollector>. The Data Collector API, can be accessed at <http://epiwork.di.fc.ul.pt/collector>.
- 3- A forum for discussion and sharing of information among users of the platform. The forum is available at <http://epiwork.di.fc.ul.pt/forum>.
- 4- The Mediator, providing RESTful webservices that can be used by foreign applications for the automatic search, download and upload of data and metadata to the repository. The mediator will be made available at <http://epiwork.di.fc.ul.pt/mediator>.

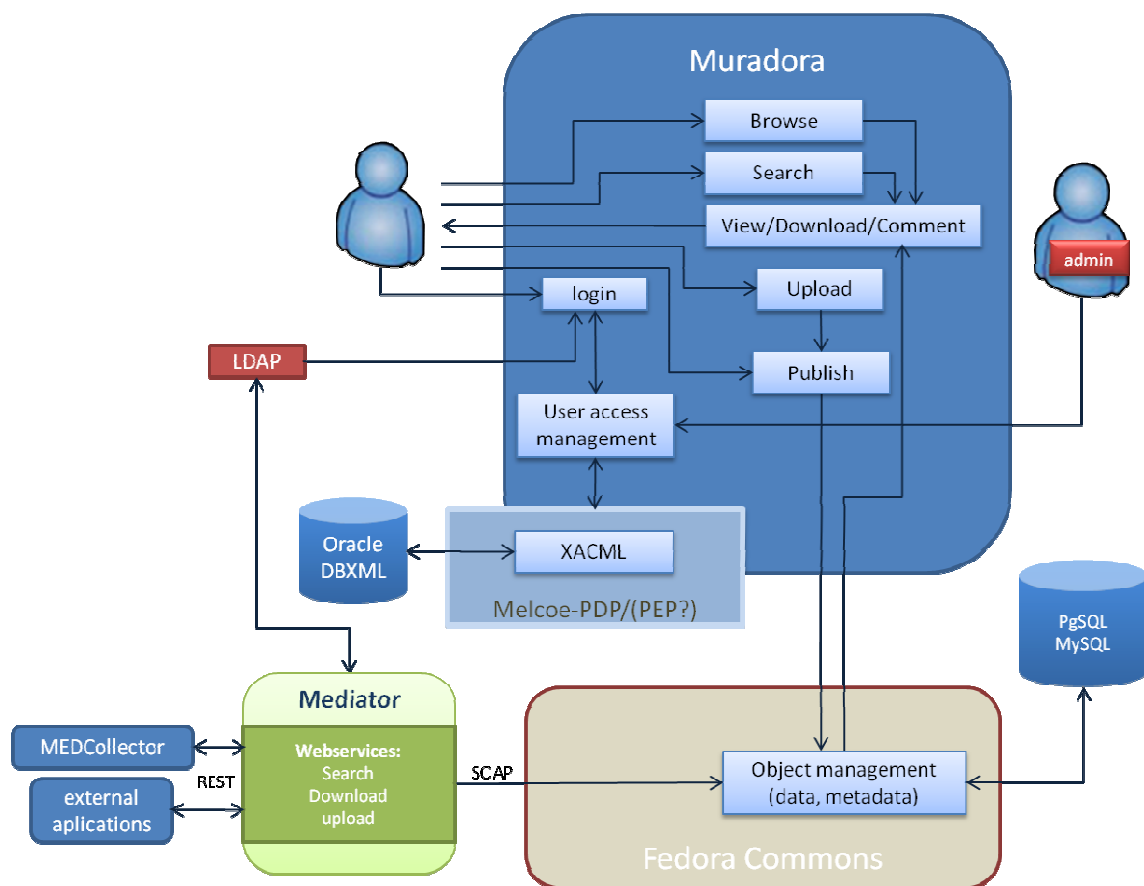


Figure 3- The digital repository and its integration with the other EM modules and external applications. The digital repository is composed by Fedora Commons and Muradora, which is the front end of the repository. It also makes use of other software such as LDAP, Oracle DBXML and SQL databases. The Mediator provides webservices which can be used by other applications to accesses/deposits data in the repository.

Figure 3 presents a scheme of the EM, excluding the Forum. The digital repository provides tools to manage resources which can be uploaded (or just referenced) and annotated, browsed, searched, commented and downloaded. Users can access the repository through a web frontend, while applications can access the repository through RESTful webservice provided by the Mediator. At the current state, the integration of the Forum with other modules is limited to the use of the same LDAP database for user login. However, it will progressively evolve to a situation where it is completely integrated with the repository frontend.

The EM repository manages not only datasets physically stored in the system, but also metadata about these datasets and about other resources that cannot be physically stored in the Epidemic Marketplace. This may be due to security or owner's rights constraints or due to their nature (for example, resources such as web tools, organizations or events can not be stored physically, though they can be referenced).

The objective of this work is the development of a metadata model for the Epidemic Marketplace.

1.2 Motivation

The first motivation for the development of a specific metadata model for the EM is to provide a more structured way to annotate data with metadata. This will be essential to describe epidemiological resources in a more standardized and accurate way and it will improve data management and search in the repository.

Moreover, as seen in the previous section, the EM is composed by different modules and is further expected to interact automatically with foreign applications. Having data correctly described using a semantically rich metadata framework is essential to improve interoperability and to allow the automatic manipulation of data.

The interaction between the different modules and with external applications will be done through the Mediator, which will expose RESTful webservices for data search, upload and download. These operations will be facilitated by the use of a more specific metadata schema.

1.3 Objective

The objective of this work was to develop a metadata model to be used in Epiwork's EM. The aim of this metadata model is to enhance epidemiological data management and manipulation.

The first goal was to design a metadata model based on metadata standards, using specific metadata elements for a relevant and consistent description of epidemiological datasets and relying on encoding schemes and controlled vocabularies for metadata standardization and semantic enrichment.

The second goal was to implement a form, based on the EM metadata model, for the annotation of epidemiological resources. The form for resource annotation with metadata annotation form which can provide support for the creation of the metadata contents, thus making this process easier and more user-friendly.

1.4 Methodology

The first step in this work was the analysis of available software to build a digital repository for the EM. Having identified the main open source options available and after discussion with the other members of the project, it was decided to build the repository with Fedora Commons, using Muradora as frontend.

The next step was the design of the metadata model. Existing standards for metadata annotation were analyzed and it was decided to use the Dublin Core (DC) standard as a base for the design of the EM metadata model. Specific metadata elements were then defined to annotate epidemiological related resources. Encoding schemes and controlled vocabularies were identified and included for metadata standardization and to add semantic meaning.

Finally, it was created an extension for the digital repository in order to provide the tools to annotate resources using the EM metadata schema. This was done by creating a XForms script, implementing mechanisms to support metadata creation.

1.5 Results

A first prototype of the EM digital repository was deployed. It provides tools to upload or reference resources and to annotate them with metadata as well as tools to search and download these resources or their metadata. The access to the digital repository is controlled according to specific user permissions defined according with user data stored in LDAP and with policies defined in XACML files.

A first version of the EM metadata model was designed and has been under constant revision and improvement. The repository now provides a form to insert and edit metadata in the EM metadata schema format. The form, to annotate resources with metadata, has suffered several revisions and includes automatic mechanisms to support the annotation with metadata.

This work has been presented at ICDL 2010 - The International Conference on Digital Libraries, which took place in 2010 at New Delhi, India. A full paper has also been accepted for the 1st International Conference on Information Technology in Bio- and Medical Informatics (ITBAM '10) at DEXA (Database and Expert Systems Applications) 2010. A journal article is also under preparation.

1.6 Organization of the document

This document is organized in six chapters, according with the following description.

Chapter 2 gives a general overview of related work on data management, with special focus on health and epidemic data management in the internet. It is introduced the concept of metadata and metadata schemas, for data annotation and management, and the use of ontologies for the semantic enrichment of metadata. Other technologies are introduced such as digital repositories and OAI (Open Archives Initiative) standards for data and metadata exchange.

In Chapter 3, the repository requirements are stated. The setup of the repository and its hardware base are also described.

In Chapter 4, the EM metadata model is presented, as well as the strategies used for its design.

Chapter 5 is about the implementation and evaluation of the metadata model in the digital repository. In this chapter are described the methods used to produce a form relying on mechanisms to support metadata creation.

In Chapter 6, it is discussed the metadata model, its relevance to the platform, conclusions and future work.

Chapter 2

Related Work on Data Management on the

Internet

Digital repositories are a great technology, making use of the internet to provide easy and controlled access to important information. In the last years, the use of digital repositories has expanded in many areas. For example, some of the most popular sites in the internet are YouTube, a digital video repository [12], and Flickr, a digital photo repository [13].

Especially in the areas of science this has brought incredible advantages to researchers who in the last year have witnessed the rise of the internet as the main provider of access to scientific literature, being possible to obtain online much of the scientific production.

For example, the Pubmed repository stores over 19 million citations (metadata), from biomedical journals and books and is currently the most used repository of bio-literature [14]. While Pubmed does not actually store the journals or books referenced, in many cases it provides links for repositories holding those resources in a digital format, making them available to the scientific community.

This was an incredible step forward, from having these resources available only in physical libraries, allowing the information to reach a much wider audience and making information much more accessible to researchers.

Many other scientific repositories have been created to store not only scientific literature but also data produced by experimental procedures. There are repositories to store genomic [15,16], protein [17,18], chemicals [19] and clinical trial data [20], just to name a few examples.

Metadata is fundamental in digital repositories to manage and organize data. Metadata is information about data, which identifies its context and content, facilitating its management and search [21].

Metadata is often used in the context of web resource annotation, though this is not a new concept. For example libraries had catalogue records about books, which are metadata, before the appearance of the internet. However, due to the huge quantity of information in the internet, and its heterogeneous structure, the concept of metadata was re-introduced there in an attempt to make information easier to be found and shared by people and machines [22,23].

However, for that to be possible, it is necessary to define a metadata schema and data needs to be annotated using controlled vocabularies [24]. The use of metadata to describe data, together with ontologies, to define concepts and relations between these concepts, is becoming common practice in information and knowledge management.

One of the driving forces for the implementation of metadata and metadata standards is the World Wide Web Consortium's (W3C) initiative for the Semantic Web [25]. The use of metadata annotated with ontologies is necessary for information to be machine-readable, which is essential for the development of the Semantic Web.

This chapter will discuss the use of metadata and ontologies in data management, especially in the area of epidemiology and health.

2.1 Metadata models

In order to obtain a useful and functional data annotation with metadata it is necessary to specify a metadata model. A model is an immaterial representation of a relevant part of the real world [26]. To design the metadata model it is necessary to analyze the data to be stored and managed, and understand how to best describe it [21].

The metadata model defines a schema for the annotation of data with metadata, including semantic definitions of the terms used in that schema, structural constraints, data structure definitions and its binding to a specific syntax such as XML (eXtensible Markup Language) [27].

The application of a well defined metadata schema standardizes metadata, enabling easier data exchange between applications, therefore improving interoperability [27]. The

implementation of metadata templates that allow automatic data manipulation, edition, and exchange is increasing, especially with the advent of semantic web [25].

In the area of health sciences some repositories have been identified using metadata schemas for the annotation of data, such as the data repository of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) [20], or the Heal repository for teaching resources for health sciences education [28].

However, the metadata schema of the NIDDK repository relies in a small number of fields that are annotated mostly with free text. This renders metadata insufficiently structured and semantically poor.

In the case of the Heal repository, it relies in a detailed metadata schema defined using selected elements from the IMS Meta-Data Version 1.2.1 specification [29], and several elements defined locally.

From these examples, the Heal repository is the one closer to what we propose for the Epidemic Marketplace. However, since the aims of these repositories have different purposes their metadata schema requirements and construction are not the same.

To the best of the author knowledge, there is not a metadata model publicly available for the specific annotation of epidemiological resources.

2.2 Metadata standards

There are several standards for the collection and management of metadata, which are general models that can be adapted to specific applications, such as the ISO/IEC 11179, the Dublin Core Metadata Element Set (DCMES) or the Metadata Object Description Schema (MODS).

The ISO/IEC 11179 is a standard for storing organizational metadata in a controlled environment, called a metadata registry [30]. It is a standard for metadata-driven exchange of data in an heterogeneous environment, based on exact definitions of data. An ISO metadata registry consists of a hierarchy of *concepts* with associated *properties*. Here, concepts are similar to classes in object-oriented programming, but without the behavioral elements.

The ISO/IEC 11179 was designed to be used in enterprise environment. Several health organizations are known to implement this standard, such as the Australian Institute of Health and Welfare - Metadata Online Registry (METeOR) [31]; the US Health

Information Knowledgebase (USHIK) [32]; and the Cancer Data Standards Repository (caDSr), developed by the US National Cancer Institute. For example, the caDSr was developed with the goal of defining a comprehensive set of standardized metadata descriptors for cancer research data, both for information collection and analysis [33].

The DCMES, or simply DC, is a vocabulary of fifteen properties used for the description of web resources [34] and has been recognized as the standard ISO 15836:2009 [35]. The DC metadata standard was developed by the Dublin Core Metadata Initiative (DCMI) [36]. Recently the DC model has been revised and these elements have been expanded, so the 15 fields defined initially are now part of a larger set of metadata vocabularies, the DCMI Metadata Terms (DCTERMS) [37].

In order not to interfere with the use of the 15 DC properties initially defined, 15 new properties were created with the same names but as sub-properties of the older set. It is possible to choose any one of these variants to use in applications, according to their specific needs. However, the use of the latest version is recommended, because it is semantically more accurate and more consistent with the notions of best practices for machine readable metadata [37].

The MODS metadata standard was developed by the Library of Congress' Network Development and MARC Standards Office, with the objective to define a single, coherent schema for describing digital objects [38]. It is written in XML and provides 19-top level elements for describing objects, and further 64 sub-elements. While being extensible, as DC, its objective was to provide a more specific core schema, such that it can provide further functionality without the need of being extended.

In this work, we have started by using the DC schema while a specific schema for the EM was being developed. The EM schema was based on the DCTERMS, keeping some of the terms defined in that schema and including new terms specific for the EM metadata schema.

2.3 Metadata registries

Metadata schema registries are formal services that provide services over metadata vocabularies to users, human and machines, enabling the publication, navigation and sharing of information about metadata. Registries store the semantics of metadata elements, maintain information about local extensions defined in specific schemas and

provide mappings to other metadata schemas [27]. This information is essential for the interoperability of different metadata schemas, which is fundamental to render information resources shareable and discoverable.

A Dublin Core Metadata registry (<http://dcmi.kc.tsukuba.ac.jp/dcregistry>) is available to promote the discovery and reuse of properties, classes, and other types of metadata terms. It provides an up-to-date source of authoritative information about DCMI metadata terms and related vocabularies [39].

2.4 Controlled vocabularies and Encoding schemes

To achieve metadata standardization and semantic enrichment it is fundamental to use encoding schemes and controlled vocabularies.

Encoding schemes allow the standardization of formats by defining in what format data should be recorded. For example, the ISO 8601 [40] defines the format in which date should be recorded. This encoding scheme is proposed by DCMI [37] and has been adopted as the W3C standard [41]. Another example is the encoding scheme for internet language defined by the BCP47 standard [42].

Besides encoding schemes, controlled vocabularies based on information organization systems, such as taxonomies, thesaurus or ontologies should be used for data standardization. A controlled vocabulary in its most simple form defines a controlled list of preferred terms [43].

A taxonomy is a collection of controlled vocabulary terms organized in a hierarchical structure and a thesaurus defines a networked collection of controlled vocabulary terms.

The term Ontology as used in computer science was derived from the philosophical term, which defines it as a systematic account of Existence. In computer science ontologies have been defined as an “explicit specification of conceptualization” [44] that can help humans and computer applications to share knowledge. Conceptualizations refer to the entities, namely: the terms, the relationships between terms and the constraints of those relationships. Specification refers to the explicit representation of the conceptualizations.

A formal ontology should contain explicit descriptions of the concepts in a given domain, which should be organized and structured according to the relationships between them [45].

The employment of controlled vocabularies is essential for the standardization of metadata content, making information finding and indexing more efficient by defining the use of preferred terms, avoiding the use of different terms for the same concept as would be expected in natural text.

There are some tools available for the standardization of metadata in the biomedical area, such as the Unified Medical Language System (UMLS). The UMLS, which defines as UMLS Knowledge Sources three distinct components: the Metathesaurus, the semantic network and the SPECIALIST lexicon [46]. The purpose of the UMLS is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health and can be used to enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information, as well as in informatics research. By design, the UMLS Knowledge Sources are multi-purpose and are not optimized for particular application.

The Open Biomedical Ontologies (OBO) Foundry is a collaborative experiment involving developers of science-based ontologies who aim to establish a set of principles for ontology development in order to create a suite of interoperable reference ontologies in the biomedical domain [47,48]. The Open Biomedical Ontologies (OBO) also provides an ontology repository, containing openly available ontologies relevant for the epidemiological area. One of these ontologies is the Infectious Disease Ontology (IDO), which is in fact a set of interoperable ontologies that together provide coverage of this domain [49].

A controlled vocabulary of geographic concepts will also be extremely important for a consistent description of the spatial coverage. Some systems provide controlled vocabularies to describe geographic concepts in an unequivocal manner, such as Yahoo! GeoPlanet [50], GeoNames [51] and the Thesaurus of Geographic Names (TGN) [52]. INSPIRE is also a relevant on-going project aiming to develop an infrastructure for spatial information in the European Community, which was recently launched by the European Commission [53]. It lays down general rules to establish an infrastructure for spatial information in Europe.

Ontologies will have an important role in integrating heterogeneous data sources by providing semantic relationships among the described objects. Furthermore, methods and services can be implemented for the alignment of the ontologies. The aligned ontologies and the annotated datasets will eventually serve as the basis for a distributed information

reference for epidemic modelers, which will help strengthen the communication and integration of the epidemiologists' community.

2.5 Open Archives Initiative

The Open Archives Initiative (OAI) develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content supported by open access movement (www.openarchives.org).

The OAI has two ongoing projects, the OAI Protocol for Metadata Harvesting (OAI-PMH) and the OAI Object Reuse and Exchange (OAI-ORE).

The OAI-PMH is useful for DC metadata exchange using web protocols. The OAI-PMH distinguishes two types of participants: data providers, that use this framework to provide metadata about their contents; and service providers, that harvest metadata and make use of it in the services they provide [54].

The OAI-ORE defines standards for describing and exchanging aggregated web resources. These aggregations may combine resources of different types into compound digital objects [55].

2.6 Digital repositories

There are a large number of health and epidemic data repositories available in the internet. These include all types of resources relevant to the area of life sciences, such as: bibliographic data, molecular biology data, epidemiological data, demographic data and so on.

The United States National Library of Medicine presents a list of resources available on the web that provide access to epidemiological data and statistics [56]. One of the services presented is the CDC Wonder that provides access to CDC's data in the public domain [57]. These resources are extremely useful tools providing access to public datasets and statistics.

MyPubliHealth is a repository of public health resources. This website uses a metadata schema to organize and manage resources making information finding easier and faster [58].

The previously referred NIDDK data repository contains clinical trial data [20] and the Heal repository stores teaching for health sciences education [28]. The Heal repository is designed to facilitate the sharing of high-quality, freely available multimedia resources located on the HEAL server or in other remote servers. HEAL also functions as a *publishing venue*, for authors to submit multimedia resources for review and publication there.

The Clinical Data Repository/Health Data Repository (CHDR), which consists of an interface between the US Department of Defense Clinical Data Repository and electronic health records with the Health Data Repository maintained by the US Department of veteran Affairs [59]. The CHDR enables the bidirectional exchange of computable outpatient pharmacy and medication allergy data and aims to enhance decision support by permitting data from those repositories to be cross-referenced for drug-drug and drug-allergy interactions.

The Propel Population Health Data Repository (PHDR) is a repository which allows to search for, and view data from a variety of population health data sources [60]. The PHDR makes metadata available and access to data itself is provided only after an application is submitted to, and accepted, by the owners of the data.

Two platforms in the area of cancer research have been under development: the Advancing Clinico-Genomic Trials on Cancer (ACGT) [61] and the cancer Biomedical Informatics Grid (caBIG) [62]. Both this systems provide an integrated management of data, providing tools not only for data storage and exchange, but also for the analysis of data.

These systems represent different philosophies and strategies for data management and exchange. Some of them only make metadata available online while others provide not only metadata, as well as the data in some cases even tools for analysis. regardless of the approach, all this systems aim to provide a better access to information and therefore improve investigation and/or medical practice. However, issues related to data sharing, such as legal, regulatory, ethical and intellectual property, need to be considered and either data is previously anonymized or data privacy policies have to be applied.

Chapter 3

Epidemic Marketplace Digital Repository

The EM Digital Repository is a repository for epidemiological datasets and other resources relevant for epidemiological studies. This repository is part of the EM platform and besides storing and managing data and metadata it must integrate with the other modules (Figure 2).

There are several functional requirements that should be met for the repository implementation:

- **Provide a stable and robust structure to store and organize datasets** – This repository should be able to maintain and secure the stored resources, providing tools to organize those resources in collections.
- **Support the sharing and management of epidemiological data sets** – Users should be able to upload datasets and annotate them with metadata, allowing the creation of a catalogue to improve information finding.
- **Support secure access to data**– Access to data must be controlled, according to specific permissions and rights. The repository must have a registration and login system to identify individual users. It must be able to manage access to specific resources according to well defined access policies, at individual resource and collection level.
- **Distributed Architecture**- The repository should be able to support the vision of the EM as a platform to be deployed in several sites in a geographically distributed architecture. The distributed architecture should provide improved data access performance, improved availability and fault-tolerance.
- **Support the creation of a virtual community for epidemic research**- The repository should support the platform role of community discussion center. It needs

to complement the Forum module, allowing the local discussion of specific resources through user commentaries.

In this chapter, the deployment of the digital repository is described.

3.1 Setup of the digital repository

Before setting up the EM digital repository was done a survey of software available to build it. The software was chosen considering the functional requirements stated previously and considering as a priority the use of stable versions of free open source software.

There are several open source software packages available to build a repository [34,63]. The main software packages considered were: Fedora Commons [55], DSpace [64] and EPrints [65].

Fedora Commons is a content management software that runs as a web service within an Apache Tomcat web server. This software supports the creation and management of digital content objects, independently of its type. Following a digital object model it is able to combine any number and variety of datastreams, with the support of a SQL database [55]. The resources may be stored locally in the repository or just referenced. Metadata and data are treated uniformly by the digital object model so any number and variety of metadata formats may be stored as datastreams, alongside content, in the digital object. Objects can contain metadata expressing any type of relationships among them. Relationship metadata is indexed and can be searched using semantic web query languages.

However, Fedora Commons does not contain a web frontend, making it necessary to be used together with other software packages. There are several options available, for that purpose, namely: Muradora [66], RODA [67], Fez [68], Elated [69], the Islandora module for Drupal [70] and PubMan with eSciDoc [71]. The independence of the frontend, and considering that there are several options available for that role, grants extra flexibility to a solution based on Fedora Commons. The downside is that this system becomes more complex and difficult to implement than the other stand alone software packages.

The Muradora package provides some features regarded as important for the EM platform: a) support the use of metadata standards; b) support to Shibboleth-based federated authentication infrastructures important for the integration with other

components, due to the distributed vision of the EM [9]; c) highly granular role-based access controls.

Muradora makes use of middleware software to provide extra functionality:

- OpenLDAP, to store user authentication information. OpenLDAP makes use of LDAP (Lightweight Directory Access Protocol), an industry open standard directory system, capable of storing information describing users, applications, files, printers, and other network resources [72]. The use of LDAP is essential for the unification of the several EM components, centralizing in a single database the user login information.

- MELCOE PDP, to manage individual access policies using policies contained in external XML files according to the OASIS XACML (eXtensible Access Control Markup Language) standard [73]. These XACML files are stored in an instance of Oracle's DBXML [74].

Another middleware component used by Muradora is Solr. Solr runs on Tomcat and is a search server based on the Lucene Java search library [75]. Solr has XML/HTTP and JSON APIs, a web administration and features, among others, hit highlighting, faceted search, caching and replication.

Considering those characteristics we decided to use Fedora Commons and Muradora for the deployment of a first version of the repository.

3.2 Repository deployment

The installation of Fedora Commons, Muradora, and other necessary software was done using the Muradora Allinone package (version 1.3.3), which available in Muradora's web site (www.muradora.org). This package contains Fedora Commons, version 2.2.2, and other software packages necessary to build the repository, including Tomcat. It also contains a shell script to guide the user through the installation process.

The installation process of the repository was not trivial, and many issues had to be solved before a fully functional repository could be made available. Some of those issues were related to: a) Installation of some of the software components. For example, the version of Oracle's DBXML contained in the pack had to be replaced with a more recent version, after several failed installation attempts; b) Declaration of variables, since these had to be declared in specific scripts and not only at operating system environment level; c) Configuration of OpenLDAP, due to its complexity and to inexperience using and

configuring this software; d) User access configuration, using XACML; e) Lack of good documentation and poor feed-back from the developer community.

After spending a considerable time solving these issues, the repository was successfully installed and made available at <https://epiwork.di.fc.ul.pt/muradora> (Figure 4).

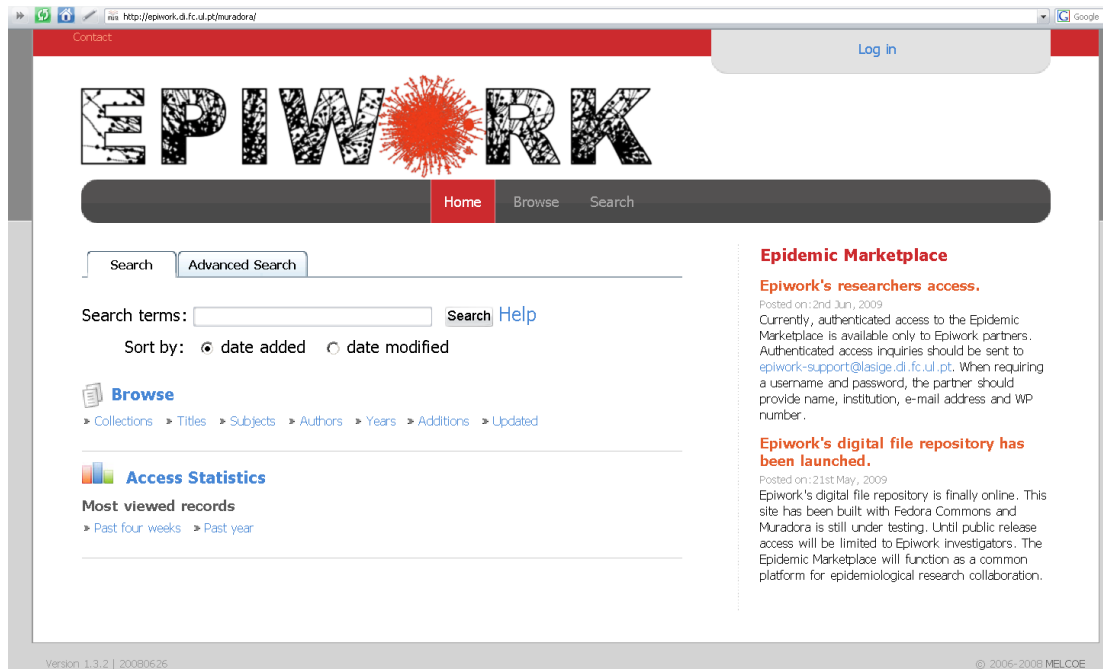


Figure 4- The EM repository main page.

3.3 Hardware setting

The repository, as well as the whole EM platform, needs to be available online to a large user community that will be able to access, insert and manipulate information contained there. The repository is required to be stable and available at all time, so it should have a stable hardware base, as well as a fast network and internet connection.

For the local deployment, the project has two DELL servers (PowerEdge SC1435), complemented with two Iomega network storage units (StorCenter pro NAS 200r1). Both servers run CentOS, a free open source Linux-based operating system.

This system provides a good level of redundancy and backup capacity, making crash recovery possible in short time. Moreover, this setting allows the use of one server for development/testing purposes while keeping a stable version publicly available.

Connectivity is provided by a 2Gb/s shared link between University of Lisbon and FCCN, which is the institution responsible for the Portuguese infrastructure of advanced

network for collaboration and communication in the fields of education and research. FCCN deploys a 10 Gb/s connection to GÉANT, the European network for research and high education [76].

Chapter 4

Metadata Model for the Epidemic Marketplace

The main objective of this work was the development of a metadata model. This model is essential for the improvement of epidemiological data management and discovery in the EM platform. The model aims to standardize and improve the communication and data exchange, both between the different EM modules and with external application.

As referred in Chapter 3, the EM digital repository was build with Fedora Commons using Muradora as frontend. Muradora provides default support to different metadata schemas, among which are the DC and MODS schemas. Moreover, it supports the extension with other metadata schemas.

In this chapter are discussed the metadata model requirements and its design, explaining how the metadata application profile is organized. It is also done a detailed description of the metadata model.

The EM metadata model design was based on the DC standard. This standard has been developed for the description of resources in the internet and was designed to be extensible so it could be applied to any area of knowledge. The EM metadata model uses terms defined by DCTERMS [37] in combination with terms defined specifically for the EM, as recommended by DCMI [77].

The EM repository will contain resources, or descriptions of resources, relevant for epidemiological/health studies. Thus, it is necessary to include specific metadata elements for an informative description of these types of resources.

Due to the interdisciplinary nature of epidemiological studies, related resources and datasets in particular may be very heterogeneous. For example, it is common to have geo-referenced epidemiological studies and so geographic datasets are expected. Other datasets could contain, for example, socio-economical, demographic or environmental data. Furthermore, data heterogeneity is increased by the use different methodologies in similar

studies. This means that a metadata schema for this repository will have to cover contents of a quite extensive and diverse area of study and nature.

A problem that may arise from this situation is the difficulty to obtain a schema that, while not too complex is capable of describing the most relevant features of resources stored in the repository. Therefore, the identification of a point of equilibrium is one of greatest challenges of this work.

This work can be complemented by the analysis of epidemiological scientific papers. This was done by characterizing data used in epidemiological papers and from there extrapolating information about datasets that may have been used in those studies.

4.1 Requirements

There are some important requirements to take into account when designing a metadata schema:

- 1- **Represent the principal concepts.** The metadata model must be able to support the description of the principal concepts represented in the resources being described. Since epidemiologists, and other health professionals, are the target users of this platform, it is necessary to understand what kind of information is more important for them. Different users might look for different information in the same dataset or have different data requirements.
- 2- **Based on standards.** The metadata schema should be based on metadata standards and metadata itself should be standardized. This can be accomplished by limiting the use of free text and by using encoding schemes and controlled vocabularies.
- 3- **Promote interoperability.** A metadata schema based on standards is essential for the interoperability between applications.
- 4- **Be user friendly.** Both the metadata schema and its implementation should consider user friendliness as an important feature. This can be achieved by having a concise metadata schema as well as by the implementation of mechanisms to support the creation of metadata.

4.2 Application profile

Application profiles contain detailed descriptions of metadata schemas, containing the requirements and instructions for its implementation. These schemas may combine data elements from one or more namespaces and are optimized for a given application [27]. The DCMI supports the use of application profiles by providing a guide for the preparation of a DC application profile (DCAP) [77]. An HTML version of the DCAP developed for the EM may be accessed at: <http://epiwork.di.fc.ul.pt/metadata/em-dcap.html>.

In the DCAP are defined the following aspects:

1) The functional requirements of the application, i.e., the purpose of the model and how it will be used in the application.

2) The types of resources described by metadata and their relationship, as well as the metadata elements to describe them.

3) Description of the metadata elements and rules to fill them. This includes the identification of which information is intended to be included in this element and whether that information is mandatory, recommended or optional;

4) The syntax and data formats for data encoding, considering that standards should be used whenever possible. For example, the international classification of diseases (ICD), actually in its tenth version (ICD-10), should be used to define the nomenclature of disease [78]. For instance, the date format can be defined following a standard format, such as YYYY-MM-DD, as defined by ISO 8601 [40]. Another example is the annotation of geographic data, which should be based in a controlled vocabulary, such as Yahoo! GeoPlanet [50], GeoNames [51] or the TGN [52].

4.3 Definition of metadata elements

The most important concepts in the data to be annotated should be described in specific metadata elements. This helps to structure metadata and to find information more easily.

In a general metadata schema, such as the one defined by the DC Element Set or by DCTERMS, there are no specific elements to describe fundamental epidemiological concepts, so it is necessary to define and include them.

One of the first steps for the metadata model development was to define the metadata elements composing the metadata schema. To do that, the metadata elements provided by DCTERMS were analyzed in order to include those that could be useful for the EM metadata schema. After defining this preliminary metadata element set, it was complemented with other relevant metadata elements.

The identification of specific metadata elements was supported by the analysis of available datasets, thus allowing the definition of the following aspects:

- 1- Which metadata elements needed further refinement to best describe the resources? For example, the **spatial** element, defined by DCTERMS, was extended in order to include more specific information about geographic coverage. The sub-elements **country**, **region** and **city** were included in order to better structure information about these geographic entities.
- 2- Whether metadata elements should be mandatory, recommended or optional and their cardinality.
- 3- New metadata elements to cover specific information, relevant for epidemiological studies that are not included in the set provided by DCTERMS.

The metadata elements used in the EM metadata schema are described in detail in Appendix A.1 and A.2.

4.4 Application profile organization

The Epiwork repository Application Profile uses terms from two namespaces:

- DCMI Metadata Terms [<http://dublincore.org/documents/dcmi-terms/>]
- EM Terms [<http://epiwork.di.fc.ul.pt/em-terms/>]

In DC metadata descriptions, all references to terms (properties, classes, data types) are made using URIs. The Qualified Names are used as abbreviation terms for URIs and are formed by a concatenation of <prefix> + ":" + <local-part> [79]. The prefix is defined in Table 1, where their associations with URIs are defined.

Table 1- Table defining the Vocabulary Title, the Namespace Name and the Prefix used to abbreviate the Namespace name.

Vocabulary Title	Namespace Name	Prefix
Dublin Core Terms	http://purl.org/dc/terms/	dcterms
EM Terms	http://epiwork.di.fc.ul.pt/em-terms/	em

Table 2, defines the structure of the tables used to describe each metadata element in Appendix A.1 and A.2.

Table 2- The description of each metadata element used in this schema will be done according to the guidelines in this table (adapted from [80]).

Name of Term	A unique token assigned to the term
EM Term URI	The Qualified Name used in the EM schema.
EM qualified name	A Uniform Resource Identifier used to identify the term in the EM schema.
DC Term URI	The Qualified Name used in the DC schema.
DC qualified name	A Uniform Resource Identifier used to identify the term in the DC schema.
Label	A human-readable label assigned to the term.
Defined By	An identifier of a namespace, pointer to a schema, or bibliographic reference for a document within which the term is defined.
DC Definition	The definition of the term in the DC namespace.
EM Definition	The definition of the term in the EM namespace.
DC Comments	Comments on the term in the DC namespace.
EM Comments	Comments on the term in the EM namespace.
Refines	The described term semantically refines the referenced term. A refinement makes the meaning of the element narrower or more specific. It will share the meaning of the unrefined element but with a more restricted scope.
Refined By	The described term is semantically refined by the referenced term.
Has Encoding Scheme	The described term is qualified by the referenced encoding scheme. Using an encoding scheme will aid in the interpretation of an element value. These schemes include controlled vocabularies and formal notations or parsing rules. A value expressed using an encoding scheme will thus be a token selected from a controlled vocabulary (e.g., a term from a classification system or set of subject headings) or a string formatted in accordance with a formal notation (e.g., "2000-01-01" as the standard expression of a date).
Obligation	Indicates whether the element is required to always or sometimes be present. In this application profile the obligation can be: mandatory (M), mandatory if applicable (MA), strongly recommended (R), strongly recommended when applicable (RA) or optional (O).
Occurrence	Indicates any limit to the repeatability of the element.

Table 3 contains a list of terms, included in the metadata model according to their original namespace.

Table 3- Index of metadata elements.

Properties in the DCTERMS namespace	abstract, bibliographicCitation, date, dateSubmitted, description, format, identifier, language, publisher, rights, rightsHolder, source, spatial, subject, temporal, title, type
Properties in the EM-TERMS Namespace	generalDescription, citation, DOI, ISBN, ISSN, pubmedID, typeOfWR, typeOfDoc, URL, version, author, authName, authOrg, authURL, organisation, orgName, orgURL, pubName, pubOrg, pubURL, country, city, region, tempFrom, tempTo, srcName, srcURL, srcDescription, epidemiological, diagnosticMethod, disease, drug, hostSp, hostGroup, pathoSp, pathoGroup, pathoStrain, vaccine, vector, demographic, environmental, geographic, socioEconomic, refCitation, refDOI, refPubmedID, copyright, disclaimer

4.5 Metadata schema description

The metadata schema translates into a XML schema in which it will be stored. This format has the advantage of being application independent and easily parsed, since XML parsers are easily available for most programming languages.

The idea of the schema was to be as simple as possible while maintaining a good capacity to structure metadata. The XML schema is presented in Figure 5.

```
<?xml version="1.0" encoding="UTF-8"?>
<em:em xmlns:em="http://epiwork.di.fc.ul.pt/metadata/">
  <em:title/>
  <em:subject/>
  <em:generalDescription>
    <em:abstract/>
    <em:citation/>
    <em:description/>
    <em:DOI/>
    <em:identifier/>
    <em:format/>
    <em:ISBN/>
    <em:ISSN/>
    <em:language/>
    <em:pubmedID/>
    <em:type/>
    <em:typeOfWR/>
    <em:typeOfDoc/>
    <em:URL/>
    <em:venue/>
    <em:version/>
  </em:generalDescription>
  <em:date/>
  <em:dateSubmitted/>
  <em:author>
    <em:authName/>
    <em:authOrg/>
    <em:authURL/>
  </em:author>
  <em:organisation>
    <em:orgName/>
    <em:orgURL/>
  </em:organisation>
  <em:publisher>
    <em:pubName/>
    <em:pubOrg/>
    <em:pubURL/>
  </em:publisher>
  <em:spatial>
    <em:country/>
    <em:city/>
    <em:region/>
  </em:spatial>
  <em:temporal>
    <em:tempFrom/>
    <em:tempTo/>
  </em:temporal>
  <em:source>
    <em:srcName/>
    <em:srcURL/>
    <em:srcDescription/>
  </em:source>
</em:em>
```

```

</em:source>
<em:epidemiological>
  <em:diagnosticMethod/>
  <em:disease/>
  <em:drug/>
  <em:hostSp/>
  <em:hostGroup/>
  <em:pathoSp/>
  <em:pathoGroup/>
  <em:pathoStrain/>
  <em:vaccine/>
  <em:vector/>
</em:epidemiological>
<em:demographic/>
<em:environmental/>
<em:geographic/>
<em:socioEconomic/>
<em:bibliographicCitation>
  <em:refCitation/>
  <em:refDOI/>
  <em:refPubmedID/>
</em:bibliographicCitation>
<em:rights>
  <em:rightsHolder/>
  <em:copyright/>
  <em:disclaimer/>
</em:rights>
</em:em>

```

Figure 5- XML of the EM metadata schema.

The structure of the XML tree was defined considering not only the metadata schema but also its implementation. Some related elements were grouped, avoiding the creation of more than two levels in the XML tree.

According with the metadata elements description, the only mandatory elements in the metadata schema are the **title**, **format** and **type**. The **subject** metadata element is defined in this schema by a controlled list of general topics locally defined.

The **generalDescription** group (Figure 5) contains several metadata elements that, such as the name suggests, provide a general description of the resources. These include the elements: **abstract**, **citation**, **description**, **DOI**, **identifier**, **format**, **ISBN**, **ISSN**, **language**, **pubmedID**, **type**, **URL**, **venue** and **version**. **Abstract** refers to the abstract of a published resource; the **citation** element is made available to include a citation of the resource (for example, if the resource is a book or a scientific paper its reference may be inserted here); **description** provides a way to include important information about the resource that could not be described in a more specific way in any other field; **DOI** stores the digital object identifier of the resource; **identifier** stores an identifying code for the resource; **format** is the MIMEtype of the resource, which only applies to files physically stored in the repository; **ISBN** to store the ISBN of the resource; **ISSN** to store the ISSN of

the resource; **language**, for which a set of languages is provided and annotated using the language name and language code according with BCP47 encoding scheme; **pubmedID** to store the Pubmed ID code; **type** defines the type of the resource and is defined by a controlled list of five terms: *datasets*, *documents*, *event*, *software* and *web* (see section 4.6); **typeOfDoc** defines the type of a document (book, article, etc.) of resources of the type “document”; **typeOfWR** defines the type of web page/tools/services provided by resources of the type “web”; **URL** defines the URL where the resource can be accessed; **venue** is specific for *event* type resources and allows the identification of the place where a event is taking place (for instance, an address); **version** defines the version of the resource.

The following metadata elements regard the date: **date** and **dateSubmitted**. **Date** refers to the date in which the resource was created while **dateSubmitted** stores the date when the resource was submitted to the repository.

The next metadata blocks (Figure 5) refer to the identification of the author(s) of the resource, whether it is a person (**author**) or an organization (**organization**), and to the identification of the person who publishes (**publisher**) the resource in the repository. The authorship is identified by the **author** metadata block, or by the **organization** metadata block, in case it is attributed to an organization rather than to people. Both these block provide metadata elements to record the name of resource creator (**authName**, **orgName**) and homepage (**authURL**, **orgURL**). In the case of the **author** metadata block it also provides the possibility to include the affiliation (**authOrg**), i.e. the organization in which the author produced the resource. In the case of the publisher metadata block, it includes similar fields to the author block, providing metadata elements for the name (**pubName**), affiliation (**pubOrg**) and home page (**pubURL**).

The resource coverage is described using the following two metadata blocks: **spatial** and **temporal** (Figure 5), regarding the geographic and time coverage, respectively. The spatial block is refined by the metadata elements **city**, **country** and **region**. The temporal block contains two metadata elements: **tempFrom** and **tempTo**, which can be used to describe a time interval (or time point, if the dates in both fields are the same).

The source of data, contained by the resource, is covered in the **source** metadata group. This group provides a metadata element to record the source name (**srcName**), a source URL (**srcURL**) and a source description (**srcDescription**).

Several elements were included to provide an advanced resource description, according with their data content. The emphasis is on the **epidemiological** block, which includes several sub-elements. Besides the **epidemiological** block, metadata elements are

also provided for the description of **demographic**, **environmental**, **geographic** and socio-economical (**socioEconomic**) resources (Figure 5).

The **epidemiological** group, for the advanced description of epidemiological data, is refined by ten sub-elements: **diagnosticMethod**, to identify a diagnostic method/procedure used; **disease**, to identify diseases covered by the resource; **drug**, which can be used to identify drugs used, for example, for the treatment of a disease or used in a clinical trial; **hostSp**, to record the species of the disease host; **hostGroup**, a more general group of hosts (for example, if the hosts belong to a wider taxonomic group or to other non taxonomic groups); **pathoSp**, to record the species of the pathogen(s) covered by the resource being described; **pathoGroup**, same as hostGroup but applying to the pathogenic agent instead of the host; **pathoStrain**, to describe the strain of the pathogen (more specific than pathoSp); **vaccine**, to identify a vaccine used in a study, such as a vaccine clinical trial; **vector**, to identify a disease vector.

The next metadata element block (Figure 5) is used to identify documents that reference the resource being described: **bibliographicCitation**. This metadata element is further refined by three sub-elements: **refCitation**, insert a citation of the document that references the resource; **refDOI**, to record its DOI; **refPubmedID**, to record its Pubmed identification number.

Finally, the last metadata element block presented in Figure 5 is the **rights** group. The rights are refined by three sub-elements: **rightsHolder**, to identify the owner of the resource; **copyright**, to include a copyright statement; **disclaimer**, to include a disclaimer about the resource.

More detailed information about these metadata elements may be consulted at Appendix A.1 and A.2.

4.6 Resource type and metadata elements

This metadata schema presented can be adapted to describe different types of resources (Figure 6), but not all the metadata elements will be used for the description of every resource. For instance, only a subset of these metadata elements will be presented during the process of annotation of a dataset, while others will be presented for the description of a document or an event. This represents different ways of using and visualizing the schema described in the previous section.

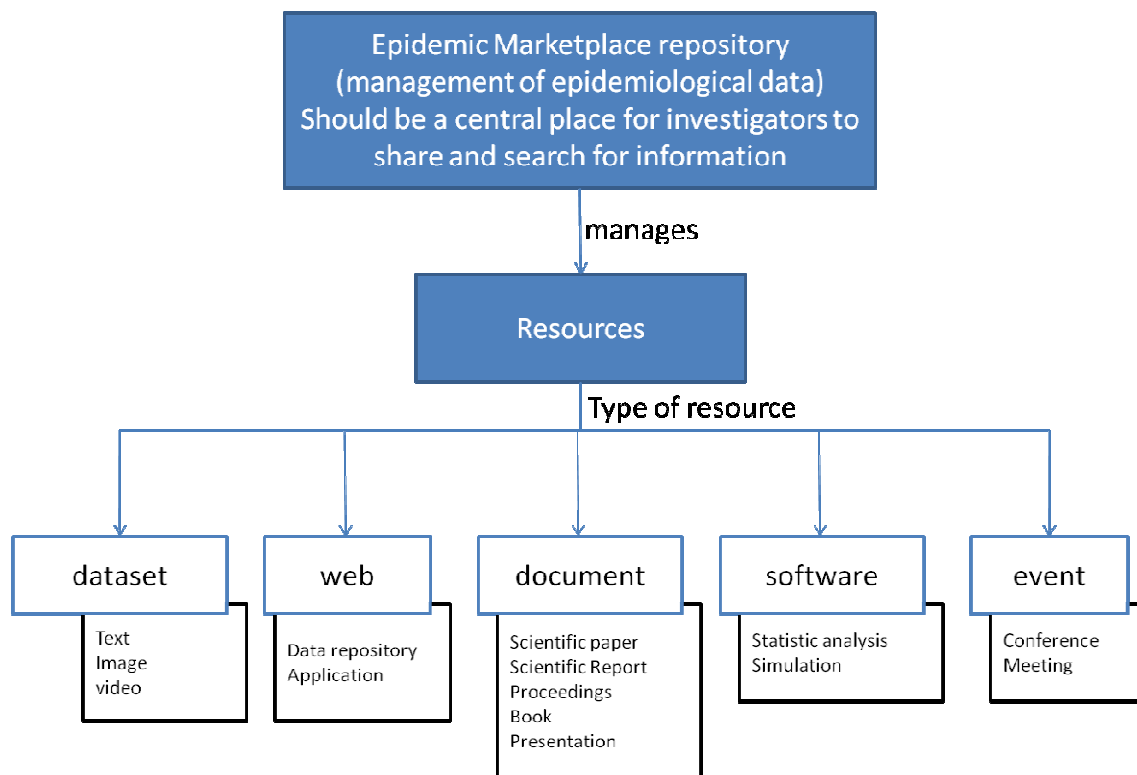


Figure 6- Types of resources managed by the EM.

The scheme in Figure 7 shows some of the metadata elements or groups of elements to be used to describe the “dataset” type of resource. This will be the main type of resource managed by the EM.

Title, **subject** and **date** are elements common to all types of resources. Other metadata elements, or groups of elements, are used in a more variable way to describe different types of resources. Figure 7 shows other metadata element groups that are important for the description of resources of the type *dataset*: **source**, **bibliographicCitation**, **rights**, **coverage**, identification and advanced description metadata blocks. Only a subset of the general description (**generalDescription**) metadata element group is used to describe a dataset: **description**, **DOI**, **format**, **identifier**, **language**, **type**, **URL** and **version**.

To describe resources of the type *document* (Figure 8), a similar set of metadata elements will be used. The main changes observed for the description of documents are at the **generalDescription** group with the inclusion of the extra metadata elements. **abstract**, **citation**, **ISBN**, **ISSN**, **pubmedID** and **typeOfDoc**.

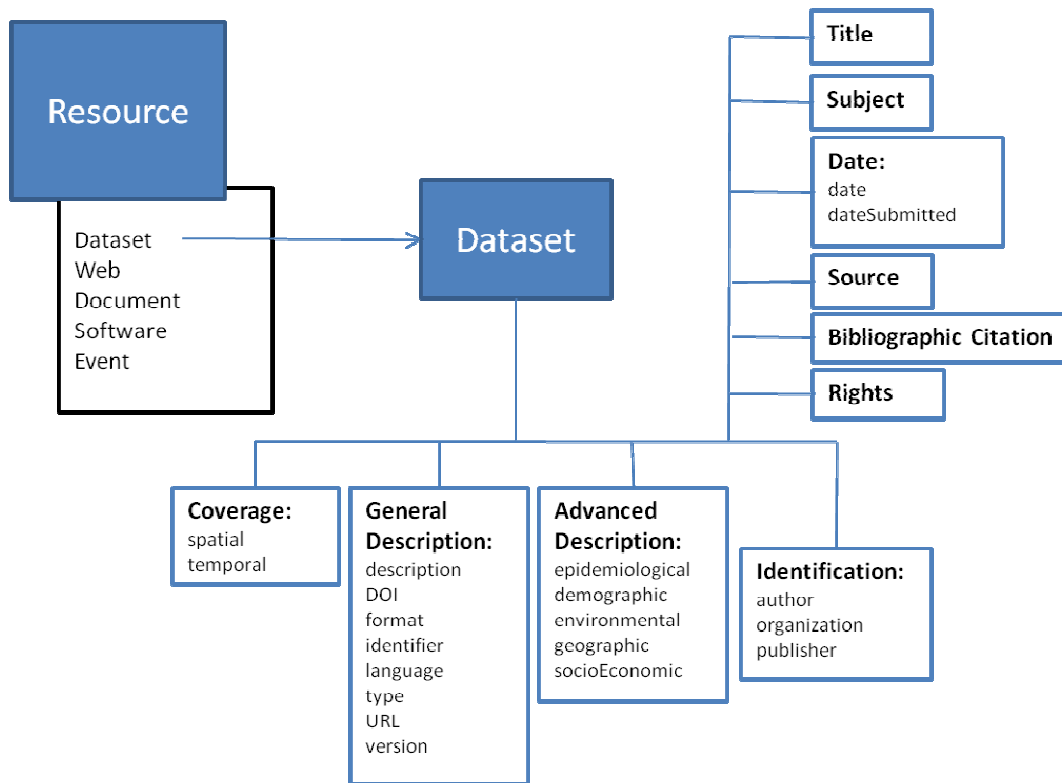


Figure 7- Metadata elements to be used to describe resources of the *dataset* type.

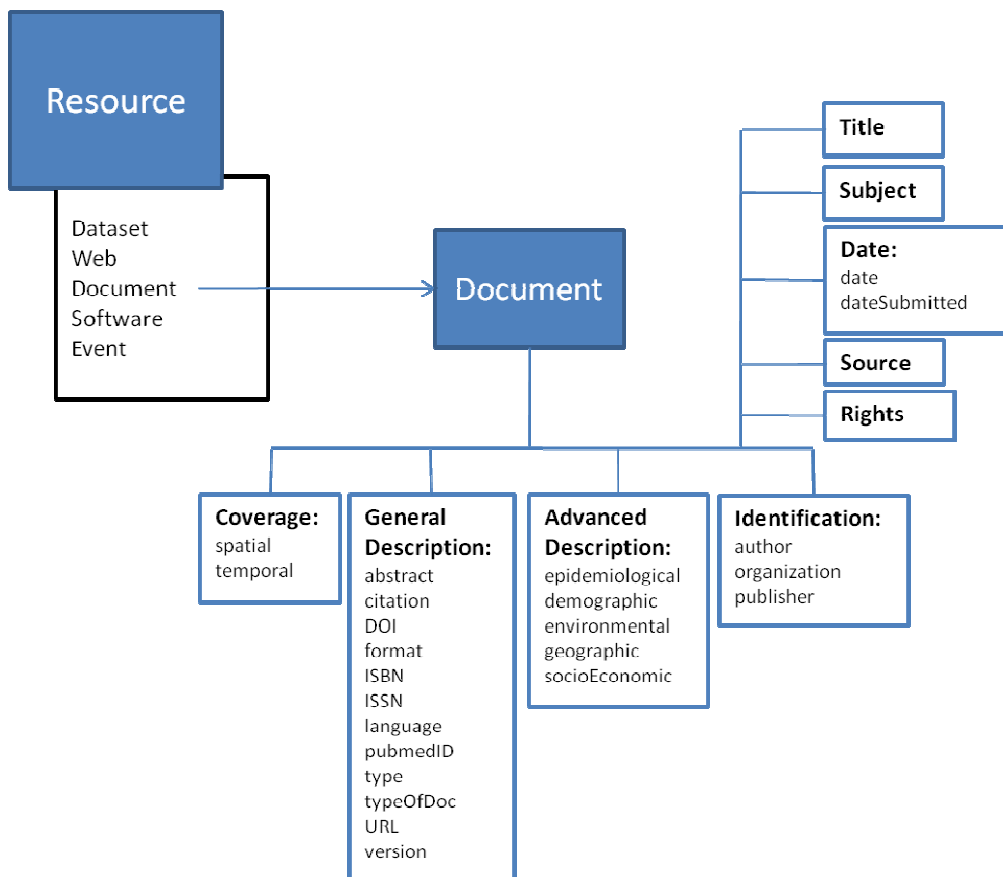


Figure 8- Metadata elements to be used to describe resources of the *document* type.

The description of resources of the type *web* also presents few differences in relation with the metadata element subset defined for *dataset* type resources (Figure 9). One of these differences is the absence of the **rights** metadata element group, the other is the absence of the **format** element, from the **generalDescription** metadata element group, since it is not stored directly in the repository and only a reference to resource is kept. On the other hand, the metadata element *typeOfWR* is included in the **generalDescription** group.

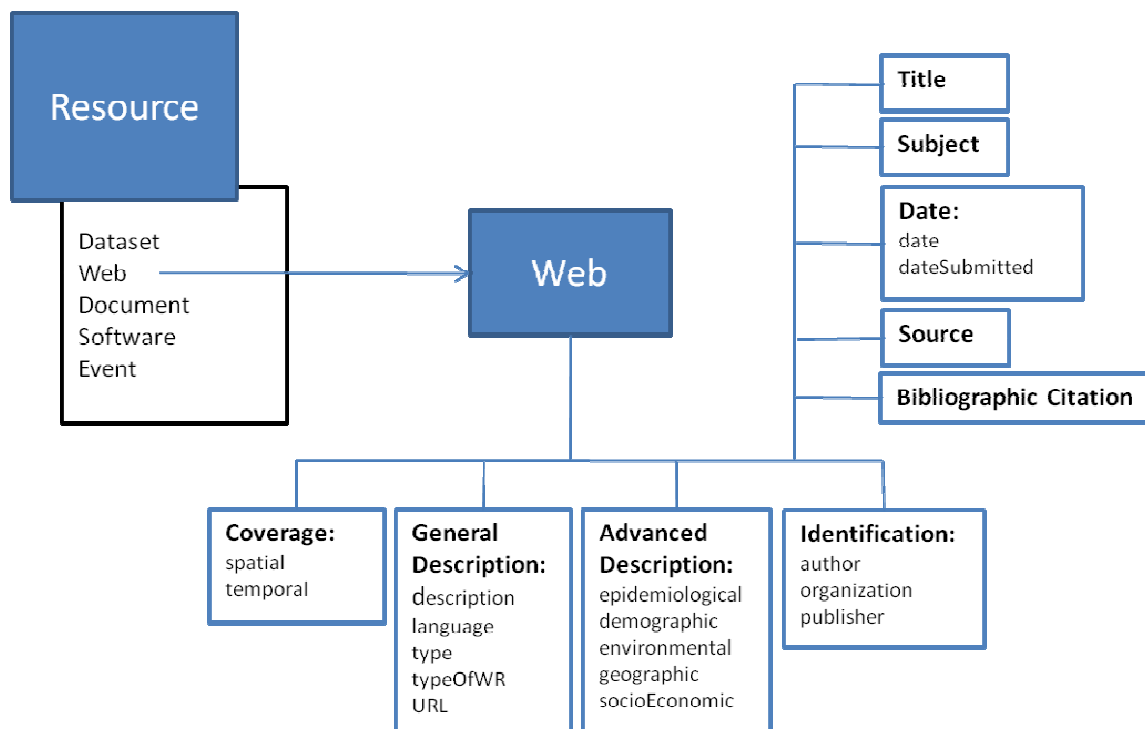


Figure 9- Metadata elements to be used to describe a resource of the type “web”.

The framework made available for the annotation of a resource of the type *event* is much simpler than for the previous types of elements (Figure 10). When comparing to the metadata profile used for the annotation of *dataset* type resources, are removed the following metadata element groups: **coverage**, all the advanced description elements, **bibliographicCitation**, **source** and **rights**. The **author** metadata block, in the identification metadata group, is also not requested. In the **generalDescription** metadata element group is included the metadata element **venue** and are removed the metadata elements: **DOI**, **identifier**, **format** and **version**.

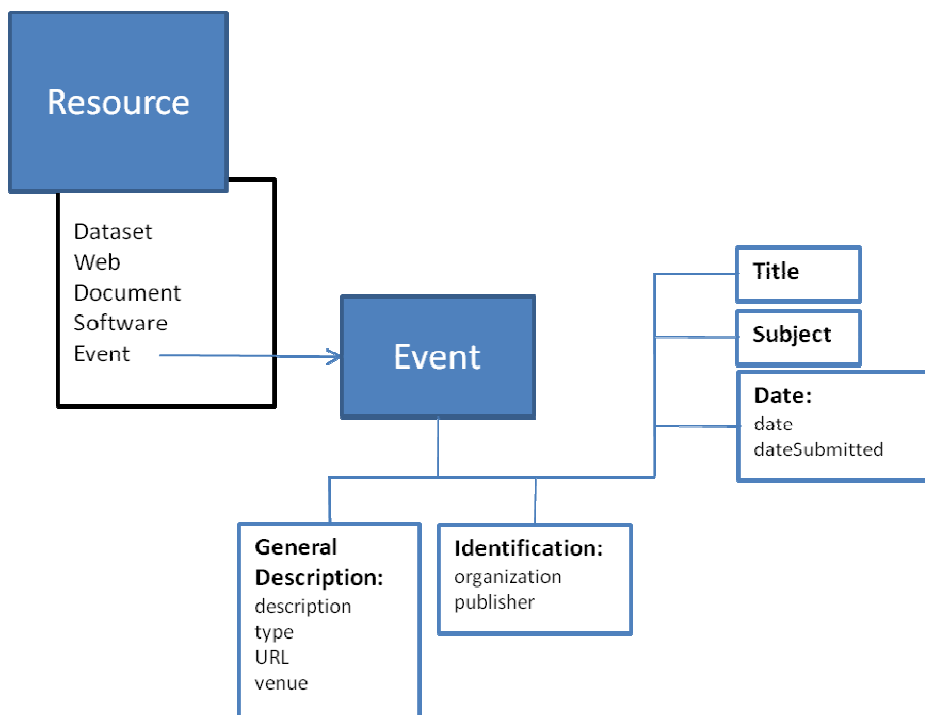


Figure 10- Metadata elements to be used to describe a resource of the type *event*.

Finally, for the description of resources of the type *software* (Figure 11), the **source** and **coverage** metadata element groups are not used, as well as all the metadata elements included in the advanced description (**epidemiological**, **demographic**, etc.).

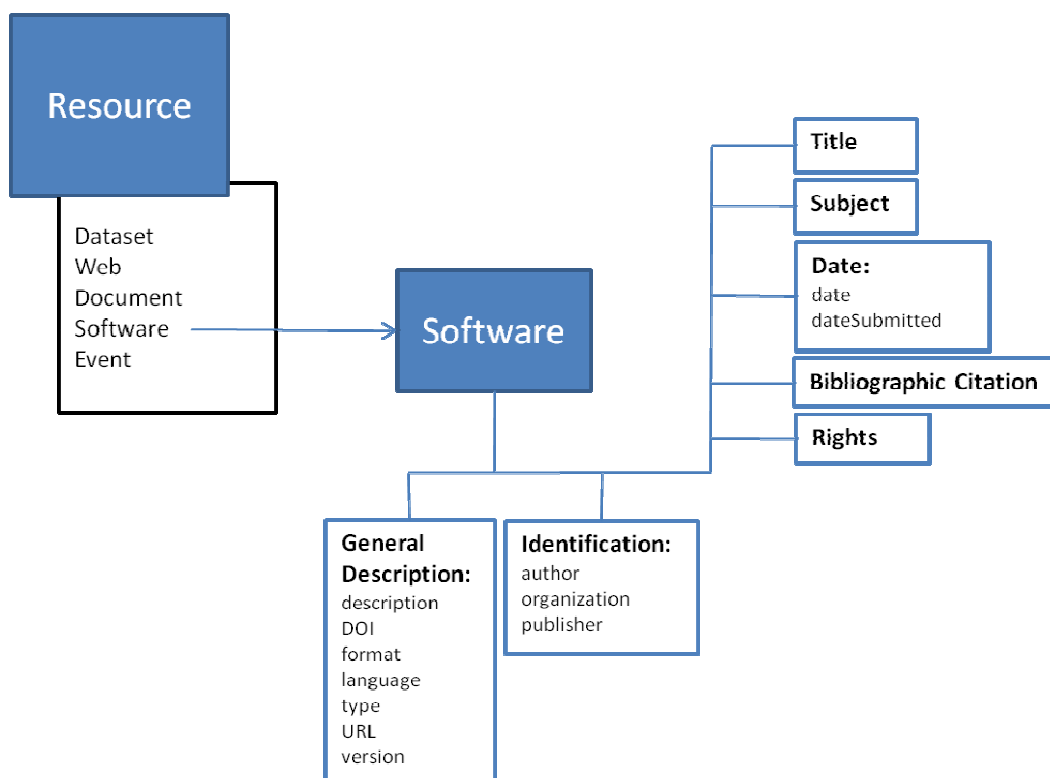


Figure 11- Metadata elements to be used to describe a resource of the type *software*.

This adjustment of subsets of metadata elements to the different types of resources helps to keep the metadata annotation process more simple and specific.

Chapter 5

Implementation and Evaluation of the EM

Metadata Model

The EM metadata model was implemented in the digital repository. This is a crucial step, since this provides an easy way for users to access, test and evaluate it. The implementation of the metadata model was done according to the instructions defined in an application profile document, following DCAP guidelines. This document contains guiding principles for the design of annotation forms, defining specific encoding schemes and controlled vocabularies to support the annotation process.

The first step in the implementation of the metadata schema was the design a new form following XForms standards. The Xforms was coded in a JSP (Java Server Pages) script, which is the coding language in which pages Muradora are built. The metadata form was complemented with help menus and automated mechanisms to facilitate the metadata filling. Controlled lists were also created for some of the metadata elements, which aim to make the filling faster and easier. Encoding schemes are being used for the annotation of the date and language.

As described in Chapter 3, the first prototype of the repository was implemented with Fedora Commons [55] using Muradora [66] as a frontend. Fedora Commons does not limit the use of metadata, since it treats metadata the same way as data, storing any kind or number of metadata annotations in their own datastreams, in the same digital object as the data described. In spite of being able to store any type of metadata profile, Fedora Commons does require the inclusion of metadata in DC format, which means that metadata from other metadata schemas need to be transformed to DC.

In Muradora, this transformation is done when the annotated resource is submitted to the repository. For that purpose, a XSL (Extensible Stylesheet Language) script is used to transform the metadata in any other format to DC, thus storing it in both formats [66].

5.1 Form design and implementation

The form for metadata annotation was based on an existing MODS metadata form. The form is divided in blocks, mirroring the XML schema organization (Figure 5). In this section is presented the form designed for the description of resources of the type “dataset”.

The first element in the form is the title field, the only mandatory metadata element that is signaled by presenting a red tone when empty (Figure 12). The form presents some aids to the user: a help menu and the automated filling of the field. The help menu can be accessed by placing the mouse cursor on the question mark symbol, such as the one in front of the word “Title”. The automated filling of the title is done using the name of the uploaded file.

The image shows a screenshot of a metadata form for a dataset description. It is organized into four distinct blocks, each with a light orange header and a light green body. The first block is titled '+ Title ?' and contains a red 'x' icon and the label 'Title' next to an empty text input field. The second block is titled '+ Subject ?' and contains a red 'x' icon and the label 'Subject:' next to a dropdown menu showing 'Epidemic'. The third block is titled '+ General Description ?' and contains a red 'x' icon, the label 'Description:' next to a large empty text area, and the label 'Language:' next to a dropdown menu showing 'Not applicable' and a question mark icon. The fourth block is titled 'Dataset creation date ?' and contains the text 'Sunday July 11, 2010' followed by a calendar icon.

Figure 12- Metadata elements required for a general description of the resource.

Help menus have been deployed for most metadata elements when deemed necessary. Whenever the symbol with the question mark is present means there is a help text associated.

The second element presented is the Subject (Figure 12). The form provides a controlled list of subject keywords created for the EM. This list will be extended and refined as more resources, covering new subjects, are stored in the repository.

The next block, named “General Description” (Figure 12) displays metadata elements from the **generalDescription** set. The metadata fields displayed here correspond to the **description** and **language** metadata elements. The “Description” field is a free text area, while the “Language” field is based on a controlled list, with language names associated with the BCP47 codes [42]. There are other metadata elements in the **generalDescription** metadata set that are not presented to the user. Some of these are not required for dataset description, such as **abstract** and **venue**, while others are automatically filled and not presented to the user, such as **type** and **format**.

The **date** is collected using a graphic date picker, which inserts date information according with the date encoding scheme defined by ISO8601 [40]. The **dateSubmitted** element is also automatically filled with the current date and is not presented to the user.

The figure displays three distinct form blocks, each with a light orange header and a light green body. Each block is titled with a plus sign, the name of the entity, and a question mark icon. The 'Author' block includes fields for Name (text input), Affiliation (dropdown menu with 'FFCUL' selected), and Homepage (text input). The 'Organisation' block includes fields for Name (dropdown menu with 'ISI' selected) and Homepage (text input). The 'Publisher' block includes fields for Name (text input), Affiliation (dropdown menu with 'FFCUL' selected), and Homepage (text input). A red 'X' icon is positioned to the left of the 'Affiliation' field in each block.

Figure 13- Metadata elements required for the identification of authors and publisher of the resource.

The identification blocks are presented in Figure 13. These are used to describe the authors, presented either as a person (**author**) or an organization (**organization**), and the

publisher. These metadata elements are automatically filled with user data stored in LDAP, which was collected in the registration process.

The coverage of a resource may be described using the “Spatial coverage” and “Temporal coverage” blocks (Figure 14). The spatial coverage may be described using 3 metadata elements from the **spatial** metadata set: **country**, **city** and **region**, as described in section 4.5. In the future the form should provide a controlled list obtained from Yahoo! GeoPlanets.

The temporal coverage, has the day as maximum resolution and may refer to a period (from one day to another), or to a time point (a specific day).

The figure displays two metadata blocks. The first block, titled '+ Spatial coverage', contains three input fields labeled 'Country:', 'City:', and 'Region:'. The second block, titled '+ Temporal coverage', contains two input fields labeled 'From:' and 'To:'. Each of these fields includes a small calendar icon and a red exclamation mark, indicating a date selection interface.

Figure 14- Metadata elements to describe the coverage.

The source of data is stored in the **source** metadata element set. In this form it is annotated in the “Information source” section (see Figure 15). The tree metadata elements in this set are annotated using freetext.

The figure shows a metadata block titled '+ Information source'. It contains three input fields: 'Name of the source:', 'Source URL:', and 'Source description:'. A red 'X' icon is visible to the left of the 'Source URL' field.

Figure 15- Metadata block to define the source of the resource.

Figure 16 shows the section of the metadata form designed to collect data for the following elements: **epidemiological** (“Epidemiological”), **demographic** (“Demographic data”), **geographic** (“Geographic data”), **socioEconomic** (“Socio-economic data”) and **environmental** (“Environmental data”).

Since the EM is a repository of epidemiological resources, the **epidemiological** element is the one which is more thoroughly extended by specific metadata elements in

order to provide means for a more detailed description (Figure 16). The epidemiological metadata block is structured in order to collect data about key features of epidemiological data.

The figure shows a series of five metadata blocks, each with a header and a list of fields:

- Epidemiological**: Fields include Disease (text), Host Species (dropdown: Human), Pathogen species (text), Pathogen strain (text), Vector (dropdown: Anopheles), Drug (text), Vaccine (text), and Diagnostic method (text).
- Demographic data**: Field is Type (dropdown: population).
- Geographic data**: Field is Type (dropdown: map).
- Socio-economic data**: Field is Type (dropdown: access to health services).
- Environmental data**: Field is Type (dropdown: temperature).

Figure 16- Metadata elements for an advanced description of resource contents.

The disease covered is probably the most important concept to be described, for epidemiological resources. Currently, the disease is defined by the annotator using free text. However, in the future this should change to a system where the disease is selected from a controlled list that should make use of disease codes based on coding systems defined by the ICD-10 [78] or the UMLS [46].

Most of the other metadata fields in the **epidemiological** section are also entered using free text. Controlled lists have been applied to the host species and vector elements. These lists are being populated with terms identified from the analysis of datasets in the repository and from external sources.

For the other metadata elements, in Figure 16, controlled lists were implemented. These controlled lists are also under development and, as with previous elements from the epidemiological set, this will be done analyzing datasets in the repository and other sources, such as scientific articles or the type of data made available by other data repositories.

The “Bibliographic Reference” block (Figure 17), corresponding to the **bibliographicReference** metadata set, describes documents that refer the resource being annotated. This section provides metadata fields to insert a reference citation (**refCitation**), a reference DOI (**refDOI**) or a reference Pubmed ID (**refPubmedID**).

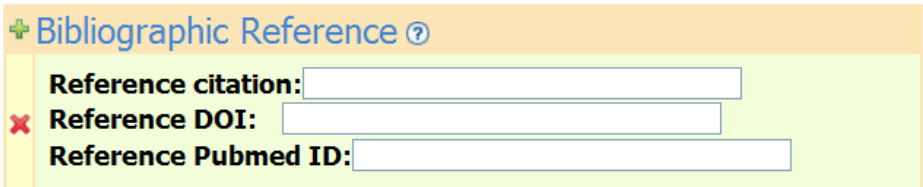


Figure 17- Group of metadata elements to describe a bibliographic reference, for the resource.

Finally, the “Rights” section, corresponding to the **rights** metadata set, provides a way to identify the owner/rights holder of the resource (Figure 18). Copyrights associated with the resources may be included here, as well as a disclaimer.

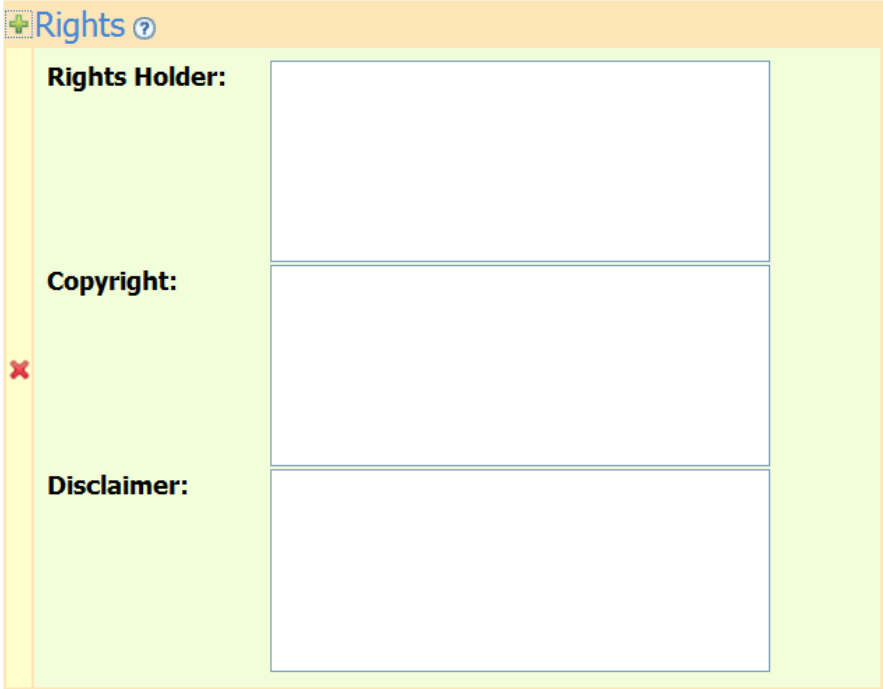


Figure 18- Metadata elements to identify the owner and include the copyright or disclaimer of the resource.

5.2 Metadata creation support

During the development of the metadata schema, it was considered as a priority to keep it as simple as possible, while still making available a structured framework for an adequate annotation of the epidemiological resources with metadata. This is important since there is a perceived resistance to entering metadata in the process of content creation [23], which may hamper the acceptance of more complex metadata schemas by the community. Therefore, it is necessary to make metadata creation as simple and straightforward as possible.

While the EM metadata model is not particularly simple, some mechanisms can lighten the metadata creation process, providing some support for the user. We propose some automatisms to support metadata creation, such as: a) assisted filling, b) auto-filling and c) metadata re-use.

To assist metadata filling some mechanisms can be used, such as help menus and dropdown menus. These speeds up the filling process by making clearer to the user what information is expected in a determined field and because selecting a topic is usually faster than writing. Moreover, it standardizes metadata since everyone will use the same term and because it avoids typing errors. Ideally, the drop down menus (or other similar aids, such as auto-complete) should present lists populated from controlled vocabularies.

Auto-filling makes metadata creation easier by inserting information automatically. For example, since the publisher in this metadata schema refers to the person who submits a resource, such as a dataset, using the user information to fill the **publisher** metadata section is certainly an advantage. In other cases, such as the **format** and **dateSubmitted** metadata elements, these are automatically filled without even presenting that information to the user, which simplifies the annotation process.

Metadata re-use is a concept yet to be implemented. The idea is that the system is able to save specific metadata profiles, thus enabling users to save and recover them when necessary. This feature presents considerable advantages when, for example, the user needs to upload a large number of similar datasets.

5.3 Testing and validating

The actual framework implemented for metadata annotation is still a work in progress. It will have to be refined and some metadata creation aids will have to be added or improved. Thus, it is important to continuously test this metadata schema, until it can be validated by the acceptance of the user community.

The metadata schema has been under constant testing, as new resources are submitted to the repository. Even though the repository is not yet open to the general public it already contains several resources submitted for testing purposes by ourselves or by Epiwork partners. It is expected that in the future the number and variability of resources stored in the repository will drastically increase, with the opening of the platform to the general public, the integration with the epidemiological modeling platform and the launch of the MEDCollector. This will provide an excellent proving ground to test and evaluate the metadata schema

When testing the metadata model, and its implementation, the key evaluation factors are:

- i) **If it represents the main concepts in the resources** – It is necessary to evaluate how much the metadata model supports an accurate and complete description of the datasets. To evaluate this it is necessary to analyze how successfully the information needed to describe the datasets maps to the metadata schema.
- ii) **If it is easy to use** – It is important to make the metadata creation process as simple as possible due to the perceived resistance of the users to insert metadata.
- iii) **If it promotes interoperability** – Being the EM composed of several different modules and aiming to automatically exchange data with external applications it is important that the metadata model helps this integration, by standardizing metadata and thus improving the automatic processing of data.

The report of the project “MultiMatch Metadata schema and mapping evaluation and revision” presents methodologies used for the evaluation of metadata schemas [81], such as:

- Analysis of the data in the resources and understand how its description maps to the metadata schema;
- Evaluation done by project partners;
- Presentation and discussion of the schema in scientific meetings.

These methods have been employed during the metadata model development. Besides constant analysis of available datasets, or datasets identified in scientific articles, the repository and the metadata schema as been presented to project partners in Epiwork meetings. Some feed-back was obtained from those meetings, mostly concerning the need to have automatic aids to keep the annotation process simple. This work has also been presented in an international meeting.

Other evaluation method to be proposed is the distribution of a questionnaire to the repository users. The feed-back obtained will hopefully be informative enough for the re-assessment and improvement of the schema, leading to the revision and improvement of the metadata model.

5.4 Analysis of datasets for the design of the metadata model

The analysis of the available datasets played an important role in the design an evaluation of the metadata model. For this purpose, were used datasets available at the EM repository as well as datasets derived from epidemiological scientific papers. This work allowed the identification of important metadata elements to be included for the description of epidemiological concepts. It also helped to understand how other metadata elements already available in the basic 15 element DC metadata schema could be extended to better structure metadata about the resource.

5.4.1 Twitter datasets

The Twitter datasets produced using an initial prototype of the Data Collector [11] and stored in the EM digital repository, contain messages (tweets) from the Twitter. The messages were collected according to their content of specific disease and location keywords. Besides the tweets, the datasets also contain for each tweet the author name (nickname), the source (in this case the Twitter.com service), the keywords searched, the date and a score.

An annotation, using the 15 element DC schema is presented in Figure 19, while the annotation of the same resource with the EM metadata schema is presented in Figure 20. From comparing both the metadata annotations it is possible to observe that the EM

metadata schema presents several advantages. It provides a much more structured metadata and allows inserting the same information using less text. Furthermore, if complemented by the use of controlled languages based on thesaurus or ontologies it makes it possible to derive much more information from there. Finally, it is much easier to be read by automatic agents.

For example, for the identification metadata fields, the EM provides extra fields in order to better structure information about the author, providing fields for author name, affiliation and homepage. While all that information could be inserted in the **creator** metadata element of the DC schema, it would be necessary to parse the text in that field to separate the information. Other metadata elements have been extended to provide this sort of functionality, such as **source**, **coverage** and **publisher**.

```
<?xml version="1.0" encoding="UTF-8"?>
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <dc:rights>Creative Commons Attribution-ShareAlike (CC BY-SA),
http://creativecommons.org/licenses/by-sa/3.0/</dc:rights>
  <dc:coverage>Spatial: TGN ID:1000090 (Portugal)</dc:coverage>
  <dc:coverage>Temporal: 2009-5-16 to 2009-6-3</dc:coverage>
  <dc:relation>Lopes LF, Zamite JM, Tavares BC, Couto FM, Silva F, Silva MJ.
Automated Social Network Epidemic Data Collector. INForum informatics symposium,
Lisbon 2009.</dc:relation>
  <dc:language>Portuguese</dc:language>
  <dc:language>English</dc:language>
  <dc:source>http://epiwork.di.fc.ul.pt/collector/</dc:source>
  <dc:identifidier>dataset-twitter-003</dc:identifidier>
  <dc:format>text/tab-separated-values</dc:format>
  <dc:type>dataset</dc:type>
  <dc:date>2009-06-04</dc:date>
  <dc:contributor>Luis F Lopes</dc:contributor>
  <dc:contributor>Joao M Zamite</dc:contributor>
  <dc:contributor>Bruno C Tavares</dc:contributor>
  <dc:contributor>Francisco M Couto</dc:contributor>
  <dc:contributor>Fabricio Silva</dc:contributor>
  <dc:contributor>Mario J Silva</dc:contributor>
  <dc:publisher>Epiwork - http://www.epiwork.eu</dc:publisher>
  <dc:title>Twitter dataset H1N1 + Portugal 4-6-2009</dc:title>
  <dc:creator>LASIGE node of the Epidemic Marketplace</dc:creator>
  <dc:subject>twitter message dataset</dc:subject>
  <dc:description>This dataset contains Twitter messages containing the words H1N1
and Portugal collected between 16-5-2009 and 3-6-2009. Information is a 7 columns
relation, containing the following data:Column 1- keyword 1 (disease)- H1N1;
Column 2- Keyword 2 (location)- Portugal; Column 3- Source (Twitter); Column 4-
Author of the message (user id); Column 5- The message body (evidence); Column 6-
score; Column 7- date (day and hour)</dc:description>
</dc:dc>
```

Figure 19- Annotation of a Data Collector dataset, using the DC metadata schema.

```

<?xml version="1.0" encoding="UTF-8"?>
<em:em xmlns:em="http://epiwork.di.fc.ul.pt/metadata/">
  <em:title>Twitter dataset H1N1 + Portugal 4-6-2009</em:title>
  <em:subject>epidemic</em:subject/>
  <em:generalDescription>
    <em:description>Data is organized in 7 columns: Column 1- keyword 1
    (disease)- H1N1; Column 2- Keyword 2 (location)- Portugal; Column 3- Source
    (Twitter); Column 4- Author of the message (user id); Column 5- The message body
    (evidence); Column 6- score; Column 7- date (day and hour)</em:description>
    <em:format>text/tab-separated-values</em:format>
    <em:type>dataset</em:type>
    <em:language>en : English</em:language>
  </em:generalDescription>
  <em:date>2009-06-04</em:date>
  <em:dateSubmitted>2009-06-04</em:dateSubmitted>
  <em:author>
    <em:authName>Luis Filipe Lopes</em:authName>
    <em:authOrg>FFCUL</em:authOrg>
    <em:authURL>http://xldb.fc.ul.pt/wiki/Luis_Filipe_Lopes</em:authURL>
  </em:author>
  <em:organisation>
    <em:orgName>LASIGE</em:orgName>
    <em:orgURL>http://lasige.di.fc.ul.pt/Main_Page</em:orgURL>
  </em:organisation>
  <em:publisher>
    <em:pubName>Luis Filipe Lopes</em:pubName>
    <em:pubOrg>FFCUL</em:pubOrg>
    <em:pubURL>http://xldb.fc.ul.pt/wiki/Luis_Filipe_Lopes</em:pubURL>
  </em:publisher>
  <em:spatial>
    <em:country>TGN ID:1000090 (Portugal)</em:country>
  </em:spatial>
  <em:temporal>
    <em:tempFrom>2009-5-16</em:tempFrom>
    <em:tempTo>2009-6-3</em:tempTo>
  </em:temporal>
  <em:epidemiological>
    <em:disease>H1N1 flu</em:disease>
  </em:epidemiological>
  <em:source>
    <em:srcName>Data Collector</em:srcName>
    <em:srcURL>http://epiwork.di.fc.ul.pt/collector/</em:srcURL>
  </em:source>
  <em:bibReference>
    <em:refCitation>Luis Filipe Lopes, João Zamite, Bruno Tavares, Francisco
    Couto, Fabrício A.B. Silva, Mário J. Silva, Automated Social Network Epidemic Data
    Collector.INForum - Simpósio de Informática September, 2009.</em:refCitation>
  </em:bibReference>
  <em:rights>
    <em:copyright>Creative Commons Attribution-ShareAlike (CC BY-SA),
    http://creativecommons.org/licenses/by-sa/3.0/</em:copyright>
  </em:rights>
</em:em>

```

Figure 20- Annotation of a Data Collector dataset, using the EM metadata schema.

Table 4- In the DC schema the disease covered by the dataset is identified in free text using non-specific metadata fields, while in the EM schema it is identified in a specific metadata field making use of a controlled vocabulary.

DC schema	EM schema
<pre><em:title>Twitter dataset H1N1 + Portugal 4-6-2009</em:title> <dc:description>This dataset contains Twitter messages containing the words H1N1 and Portugal collected between 16-5-2009 and 3-6-2009. Information is a 7 columns relation, containing the following data:Column 1- keyword 1 (disease)- H1N1; Column 2- Keyword 2 (location)- Portugal; Column 3- Source (Twitter); Column 4- Author of the message (user id); Column 5- The message body (evidence); Column 6- score; Column 7- date (day and hour)</dc:description></pre>	<pre><em:epidemiological> <em:disease>H1N1 flu</em:disease> </em:epidemiological></pre>

An example of how the EM schema is more readily read by automatic agents can be exemplified in the case where an automated agent is looking for datasets about a specific disease. In the case of the metadata annotation with the DC schema, to find the disease covered by the dataset, it would have to look for it in the **title** or in the **description**. In the EM schema an automated agent would go straight to the **disease** metadata element and collect the information there (see Table 4).

There is also the case where the free-text description may be more useful to human users. In that case the description can be included in the EM schema exactly as it is in the DC schema, since both schemas include the metadata element **description**. In this example, the content of the **description** in the EM schema is the same as in the DC schema (see Figure 20).

Table 5- Spatial and temporal coverage annotated using the DC and the EM metadata schemas.

DC schema	EM schema
<pre><dc:coverage>Spatial: TGN ID:1000090 (Portugal) </dc:coverage> <dc:coverage>Temporal: 2009-5-16 to 2009-6-3 </dc:coverage></pre>	<pre><em:spatial> <em:country> TGN ID:1000090 (Portugal) </em:country> </em:spatial> <em:temporal> <em:tempFrom>2009-5-16</em:tempFrom> <em:tempTo>2009-6-3</em:tempTo> </em:temporal></pre>

The same can be observed for the **coverage**, where the DC annotation would have to be parsed, to understand if it refers to geographic or temporal coverage that would not be the case in the annotation with the EM schema (Table 5). In this case, the **spatial** and **temporal** keywords have been included in the annotation with the DC schema, which would make the parsing easier. But this would not be the case in other situations, since the information inserted there would depend largely on the user annotating the resource. In the EM schema the spatial and temporal coverage are even further structured.

5.4.2 US Airports Dataset

Another dataset stored in the EM repository contains information about the US transportation network, namely data relative to the 500 US airports with most traffic. The dataset contains an anonymized 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.

The image in Figure 21 shows the annotation using the DC schema and Figure 22 shows the annotation with the EM schema.

```
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <dc:format>text/plain</dc:format>
  <dc:type>Dataset</dc:type>
  <dc:date>2009-09-03</dc:date>
  <dc:title>US Air Transportation Network</dc:title>
  <dc:creator>Daniela Paolotti</dc:creator>
  <dc:language>not applicable</dc:language>
  <dc:subject>Undirected weighted network of the 500 US airports with the
largest amount of traffic</dc:subject>
  <dc:description>Undirected weighted network as obtained by considering the 500
US airports with the largest amount of traffic from publicly available data.
Nodes represent US airports and edges represent air travel connections among
them. The file reports the anonymized 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.</dc:description>
</dc:dc>
```

Figure 21- Annotation of a US airport dataset, using the DC metadata schema.

The metadata about this resource is scarce so the advantage of the EM metadata schema is not so evident. Though it is clear from comparing both annotations the advantages referred before in the **coverage** and in the identification metadata. Moreover, it

is possible to identify specific geographic data in a more concise way using the EM schema. In this case since the dataset provides data about people movement in the US airports it is possible to identify this simply by adding the *transport network* and *mobility* keywords to **geographic** metadata fields (Table 6).

```
<?xml version="1.0" encoding="UTF-8"?>
<em:em xmlns:em="http://epiwork.di.fc.ul.pt/metadata/">
  <em:title>US Air Transportation Network</em:title>
  <em:subject>mobility</em:subject>
  <em:generalDescription>
    <em:description>Undirected weighted network as obtained by considering
the 500 US airports with the largest amount of traffic from publicly available
data. Nodes represent US airports and edges represent air travel connections
among them. The file reports the anonymized 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.</em:description>
    <em:format>text/plain</em:format>
    <em:type>dataset</em:type>
    <em:language>not applicable</em:language>
  </em:generalDescription>
  <em:dateSubmitted>2009-09-03</em:dateSubmitted>
  <em:author>
    <em:authName>Daniela Paolotti</em:authName>
    <em:authOrg>ISI</em:authOrg>
  </em:author>
  <em:organisation>
    <em:orgName>ISI</em:orgName>
    <em:orgURL>http://www.isi.it</em:orgURL>
  </em:organisation>
  <em:publisher>
    <em:pubName>Daniela Paolotti</em:pubName>
    <em:pubOrg>ISI</em:pubOrg>
    <em:pubURL>http://www.isi.it</em:pubURL>
  </em:publisher>
  <em:spatial>
    <em:country>TGN ID:7012149 (United States)</em:country>
  </em:spatial>
  <em:geographic>mobility</em:geographic>
  <em:geographic>Transport network</em:geographic>
</em:em>
```

Figure 22- Annotation of a US airport dataset, using the EM metadata schema.

Upon submission of this dataset many metadata elements were left unfilled, which confirms the known reluctance of users to create metadata content. Some of the metadata presented included in these annotation examples was inserted by us afterwards to facilitate the analysis.

Table 6- How the information about people mobility can be described in the DC and the EM schemas.

DC schema	EM schema
<p><em:description>Undirected weighted network as obtained by considering the 500 US airports with the largest amount of traffic from publicly available data. Nodes represent US airports and edges represent air travel connections among them. The file reports the anonymized 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.</em:description></p>	<pre><em:geographic>mobility</em:geographic> <em:geographic>Transport network </em:geographic></pre>

5.4.3 Article by Cohen and coworkers

In this study the authors correlated the risk levels of household malaria with topography related humidity [82].

Analyzing this article, we have indentified at least four distinct datasets:

1. One that contains geographic data, such as topological maps;
2. One that contains demographic data, such as mortality and birth rates;
3. An environmental dataset, containing humidity data for locations;
4. A clinical/epidemiological dataset, containing information about the population health status, pathogen, vector, diagnostics, etc.

For this example it was considered that all the data was treated together in a single composed dataset. Figure 23, shows a metadata annotation using the DC schema and Figure 24 the annotation of the same resource using the EM schema.

In the annotation process some fields were left unfilled since we had not access to the data. However, the intention of this exercise was to identify the type of data contained in the dataset and understand how it could be described in a more structured way.

Comparing both annotations, even not having access to the actual data, it is possible to identify types of data that are much better described in the EM metadata schema. Such data types can be described using keywords based on controlled languages using the following EM metadata elements: **epidemiological**, **geographic**, **socioEconomic**,

demographic and **environmental**. Otherwise, in the DC schema that information is annotation as free text in the **description** field (Table 7).

```
<?xml version="1.0" encoding="UTF-8"?>
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <dc:rights> </dc:rights>
  <dc:coverage>TGN-ID: 1000169 (Kenya)</dc:coverage>
  <dc:coverage> Western Highlands - Nandi District</dc:coverage>
  <dc:relation>J.M. Cohen, K.C. Ernst, K.A. Lindblade, J.M. Vulule, C.C.
John, and M.L. Wilson, "Topography-derived wetness indices are
associated with household-level malaria risk in two communities in the
western Kenyan highlands.," Malaria journal, vol. 7, 2008, p.
40.</dc:relation>
  <dc:type>dataset</em:type>
  <dc:language>en : English<em:language/>
  <dc:creator>Cohen, J.M.</dc:creator>
  <dc:contributor>Ernst, K.C.</dc:contributor>
  <dc:contributor>Lindblade, K.A.</dc:contributor>
  <dc:contributor>Vulule, J.M.</dc:contributor>
  <dc:contributor>John, C.C.</dc:contributor>
  <dc:contributor>Wilson, M.L.</dc:contributor>
  <dc:title> Datasets used in a study relating wetness with malaria
risk.</dc:title>
  <dc:subject>epidemiological</dc:subject>
  <dc:subject>geographic</dc:subject>
  <dc:subject>environmental</dc:subject>
  <dc:subject>clinical</dc:subject>
  <dc:subject>demographic</dc:subject>
  <dc:subject>socio-economical</dc:subject>
  <dc:description>This dataset contains demographic information about
the studied populations: birth and death rates, migration and population
size. The population economical activities were also assessed.
Environmental factors were measured: humidity and temperature. The
disease studied is falciparum malaria which is transmitted by Anopheles
mosquitoes. Household disease risk was correlated with topology and
humidity, for which house distribution was recorded.</dc:description>
</dc:dc>
```

Figure 23- Metadata annotation of a dataset, identified in the paper by Cohen and coworkers, using the DC metadata schema.

```

<?xml version="1.0" encoding="UTF-8"?>
<em:em xmlns:em="http://epiwork.di.fc.ul.pt/metadata/">
  <em:title>Datasets used in a study relating wetness with malaria
risk.</em:title>
  <em:subject>epidemiological</em:subject>
  <em:subject>geographic</em:subject>
  <em:subject>environmental</em:subject>
  <em:subject>clinical</em:subject>
  <em:subject>demographic</em:subject>
  <em:subject>socio-economical</em:subject>
  <em:generalDescription>
    <em:type>Datatset</em:type>
    <em:language>en : English</em:language/>
  </em:generalDescription>
  <em:author>
    <em:authName>Cohen</em:authName>
    <em:authOrg>Department of Epidemiology, School of public health,
University of Michigan</em:authOrg>
    <em:authURL>http://www.cshor.org/?p=68</em:authURL>
  </em:author>
  <em:spatial>
    <em:country>TGN-ID: 1000169 (Kenya)</em:country>
    <em:region>Western Highlands - Nandi District</em:region>
  </em:spatial>
  <em:demographic>Births</em:demographic>
  <em:demographic>Deaths</em:demographic>
  <em:demographic>Migration</em:demographic>
  <em:demographic>Population size</em:demographic>
  <em:environmental>humidity</em:environmental>
  <em:environmental>temperature</em:environmental>
  <em:epidemiological>
    <em:diagnostic>ICD-9-CM Procedure 90.05</em:diagnostic>
    <em:disease>ICD-10-B50 (Plasmodium falciparum
malaria)</em:disease>
    <em:hostSp>Human</em:hostSp>
    <em:pathoSp>Plasmodium falciparum</em:pathoSp>
    <em:pathoGroup>Apicomplexa</em:pathoGroup>
    <em:vector>Anopheles</em:vector>
  </em:epidemiological>
  <em:geographic>Elevation map</em:geographic>
  <em:geographic>Coordinates</em:geographic>
  <em:geographic>Altitude</em:geographic>
  <em:geographic>House distribution (coordinates/elevation)
</em:geographic>
  <em:socioEconomic>Economical activities</em:socioEconomic>
  <em:bibReference>
    <em:refCitation> J.M. Cohen, K.C. Ernst, K.A. Lindblade, J.M.
Vulule, C.C. John, and M.L. Wilson, "Topography-derived wetness indices
are associated with household-level malaria risk in two communities in
the western Kenyan highlands.," Malaria journal, vol. 7, 2008, p.
40.</em:refCitation>
    <em:refDOI>10.1186/1475-2875-7-40</em:refDOI>
  </em:bibReference>
</em:em>

```

Figure 24- Metadata annotation of a dataset, identified in the paper by Cohen and coworkers, using the EM metadata schema.

Table 7- Description of demographic, environmental, epidemiological, geographical and socio-economical data in the DC and EM schemas.

DC schema	EM schema
<p><dc:description>This dataset contains demographic information about the studied populations: birth and death rates, migration and population size. The population economical activities were also assessed. Environmental factors were measured: humidity and temperature. The disease studied is falciparum malaria which is transmitted by Anopheles mosquitoes. Household disease risk was correlated with topology and humidity, for which house distribution was recorded.</dc:description></p>	<pre> <em:demographic>Births</em:demographic> <em:demographic>Deaths</em:demographic> <em:demographic>Migration</em:demographic> <em:demographic>Population size </em:demographic> <em:environmental>humidity</em:environmental> <em:environmental>temperature </em:environmental> <em:epidemiological> <em:diagnostic>ICD-9-CM Procedure 90.05 </em:diagnostic> <em:disease>ICD-10-B50(Plasmodium falciparum malaria)</em:disease> <em:hostSp>Human</em:hostSp> <em:pathoSp>Plasmodium falciparum </em:pathoSp> <em:pathoGroup>Apicomplexa</em:pathoGroup> <em:vector>Anopheles</em:vector> </em:epidemiological> <em:geographic>Elevation map</em:geographic> <em:geographic>Coordinates</em:geographic> <em:geographic>Altitude</em:geographic> <em:geographic>House distribution (coordinates/elevation) </em:geographic> <em:socioEconomic>Economical activities </em:socioEconomic> </pre>

5.4.4 Article by East and coworkers

In this study the authors analyzed patterns of bird migration in order to identify areas in Australia where the risk of avian influenza transmission from migrating birds is higher [83]. Several datasets were identified in this study:

1. Geographic datasets, containing maps, aerial photos, addresses, etc;
2. Epidemiological datasets, comprising data from the analysis of bird infections;
3. Bird demographic datasets, with bird densities and distribution.

In this example all the data was treated as a single composed dataset. Figure 25, shows a metadata annotation using the DC schema and Figure 26 the annotation of the same resource using the EM schema.

```

<?xml version="1.0" encoding="UTF-8"?>
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <dc:coverage>TGN-ID: 7000490 (Australia)</dc:coverage>
  <dc:coverage>Wetlands and shorebird areas</dc:coverage>
  <dc:relation>I.J. East, S. Hamilton, and G. Garner, "Identifying areas of
Australia at risk of H5N1 avian influenza infection from exposure to
migratory birds: a spatial analysis.," Geospatial health, vol. 2, 2008, pp.
203-13.</dc:relation>
  <dc:type>dataset</em:type>
  <dc:language>en : English<em:language/>
  <dc:contributor>Hamilton, S.</dc:contributor>
  <dc:contributor>Garner, G.</dc:contributor>
  <dc:title>Datasets used to study waterfowl bird migration and avian
influenza transmission.</dc:title>
  <dc:creator>I.J. East, </dc:creator>
  <dc:subject>epidemiological</dc:subject>
  <dc:subject>geographic</dc:subject>
  <dc:subject>environmental</dc:subject>
  <dc:subject>demographic</dc:subject>
  <dc:description>This dataset is about the Influenza virus A, H5N1 strain,
causing Avian influenza in waterfowl birds, such as Arenaria interpres. In
this study is analysed the relation of this disease with bird population
distribution and bird migration.</dc:description>
</dc:dc>

```

Figure 25- Metadata annotation of a dataset, identified in the paper by East and coworkers, using the DC metadata schema.

This example, once again displays the better ability to structure metadata and base descriptions on keywords based on controlled vocabularies. As seen in Table 8, the EM metadata schema allows a much more structured and concise description of demographic, epidemiological and geographic data.

```

<?xml version="1.0" encoding="UTF-8"?>
<em:em xmlns:em="http://epiwork.di.fc.ul.pt/metadata/">
  <em:title>Datasets used to study waterfowl bird migration and avian
influenza transmission.</em:title>
  <em:subject>epidemic</em:subject>
  <em:subject>geographic</em:subject>
  <em:subject>demographic</em:subject>
  <em:subject>environmental</em:subject>
  <em:generalDescription>
    <em:type>dataset</em:type>
    <em:language>en : English</em:language/>
  </em:generalDescription>
  <em:author>
    <em:authName>East</em:authName>
    <em:authName>Hamilton</em:authName>
    <em:authName>Gardner</em:authName>
  </em:author>
  <em:spatial>
    <em:country>TGN-ID: 7000490 (Australia)</em:country>
    <em:region>Wetlands and shorebird areas</em:region>
  </em:spatial>
  <em:epidemiological>
    <em:disease>Avian Influenza</em:disease>
    <em:hostSp>Arenaria interpres</em:hostSp>
    <em:hostGroup>Waterfowl birds</em:hostGroup>
    <em:pathoSp>Influenza virus A</em:pathoSp>
    <em:pathoGroup>Orthomyxoviridae</em:pathoGroup>
    <em:strain>H5N1</em:strain>
  </em:epidemiological>
  <em:demographic>Migration</em:demographic>
  <em:demographic>Population distribution</em:demographic>
  <em:geographic>Map</em:geographic>
  <em:bibReference>
    <em:refCitation> I.J. East, S. Hamilton, and G. Garner, "Identifying
areas of Australia at risk of H5N1 avian influenza infection from exposure to
migratory birds: a spatial analysis.," Geospatial health, vol. 2, 2008, pp.
203-13.</em:refCitation>
  </em:bibReference>
</em:em>

```

Figure 26- Metadata annotation of a dataset, identified in the paper by East and coworkers, using the EM metadata schema.

Table 8- Description of demographic, epidemiological and geographical data in the DC and EM schemas.

DC schema	EM schema
<pre> <dc:description>This dataset is about the Influenza virus A, H5N1 strain, causing Avian influenza in waterfowl birds, such as Arenaria interpres. In this study is analysed the relation of this disease with bird population distribution and bird migration.</dc:description> </pre>	<pre> <em:epidemiological> <em:disease>Avian Influenza</em:disease> <em:hostSp>Arenaria interpres</em:hostSp> <em:hostGroup>Waterfowl birds</em:hostGroup> <em:pathoSp>Influenza virus A</em:pathoSp> <em:pathoGroup>Orthomyxoviridae </em:pathoGroup> <em:strain>H5N1</em:strain> </em:epidemiological> <em:demographic>Migration</em:demographic> <em:demographic>Population distribution </em:demographic> <em:geographic>Map</em:geographic> </pre>

5.4.5 Article by Eubank and coworkers

This article presents a system to model the spread of the variola virus in an urban setting, using bipartite graphs to model physical contact patterns that result from individual movements [84]. This article presents results for the city of Portland, in the United States of America, obtained by a system using large-scale individual based urban traffic simulations built on information from census, land-use and population mobility data.

From the analysis of this article it was possible to assess the use of several types of data:

1. Epidemiological datasets, including parameters of disease transmission.
2. Geographical datasets, including data about land-use and of transport systems networks.
3. Socio-economic data, to preview population hubs, together with land-use and transport network data.

In this example all the data was treated as a single composed dataset. Figure 27 shows a metadata annotation using the DC schema and Figure 28 the annotation of the same resource using the EM schema.

As in the previous examples, the advantage of the EM metadata schema is better perceived in its capacity to produce a structured description of specific data, such as: epidemiological, geographical and socio-economical data (Table 9).

Table 9- Description of epidemiological, geographical and socio-economical data using the DC and EM schemas.

DC schema	EM schema
<pre><dc:description> This work models the smallpox transmission among the human population of the city of Portland. This model relies on data about social contact patterns, land use and transport networks and people movement.</dc:description></pre>	<pre><em:epidemiological> <em:disease>ICD-10-B03 (Smallpox) </em:disease> <em:hostSp>Human</em:hostSp> <em:pathoSp>Variola virus</em:pathoSp> <em:pathoGroup>Poxviridae</em:pathoGroup> </em:epidemiological> <em:geographic>Transport network </em:geographic> <em:geographic>Movement</em:geographic> <em:socioEconomic>Social contact network </em:socioEconomic> <em:socioEconomic>Land use</em:socioEconomic></pre>

```

<?xml version="1.0" encoding="UTF-8"?>
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <dc:coverage>TGN-ID: 7012149 (United States)</dc:coverage>
  <dc:coverage>TGN-ID: 7014273 (Portland)</dc:coverage>
  <dc:coverage>TGN-ID: 7007708 (Oregon)</dc:coverage>
  <dc:relation>S. Eubank, H. Guclu, V.S. Kumar, M.V. Marathe, A. Srinivasan,
Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in realistic urban
social networks.," Nature, vol. 429, 2004, pp. 180-4.</dc:relation>
  <dc:type>dataset</em:type>
  <dc:language>en : English<em:language/>
  <dc:date></dc:date>
  <dc:creator>Eubank, S.</dc:creator>
  <dc:contributor>Guclu, H.</dc:contributor>
  <dc:contributor>Kumar , V.S.</dc:contributor>
  <dc:contributor>Marathe, M.V.</dc:contributor>
  <dc:contributor>Srinivasan, A.</dc:contributor>
  <dc:contributor>Toroczkai, Z.</dc:contributor>
  <dc:contributor>Wang, N.</dc:contributor>
  <dc:title>Dataset used for modeling disease outbreaks in realistic urban
social networks.</dc:title>
  <dc:subject>epidemiological</dc:subject>
  <dc:subject>geographic</dc:subject>
  <dc:subject>socio-economical</dc:subject>
  <dc:description>This work models the smallpox transmission among the human
population of the city of Portland. This model relies on data about social
contact patterns, land use and transport networks and people movement.
</dc:description>
</dc:dc>

```

Figure 27- Metadata annotation of a dataset, identified in the paper by Eubank and coworkers, using the DC metadata schema.

```

<?xml version="1.0" encoding="UTF-8"?>
<em:em xmlns:em="http://epiwork.di.fc.ul.pt/metadata/">
  <em:title>Dataset used for modeling disease outbreaks in realistic urban
social networks.</em:title>
  <em:subject>epidemic</em:subject>
  <em:subject>geographic</em:subject>
  <em:subject>socio-economic</em:subject>
  <em:generalDescription>
    <em:type>dataset</em:type>
    <em:language>en : English</em:language/>
  </em:generalDescription>
  <em:author>
    <em:authName>Eubank, S.</em:authName>
    <em:authName>Guclu, H.</em:authName>
    <em:authName>Kumar , V.S.</em:authName>
    <em:authName>Marathe, M.V.</em:authName>
    <em:authName>Srinivasan, A.</em:authName>
    <em:authName>Toroczkai, Z.</em:authName>
    <em:authName>Wang, N.</em:authName>
  </em:author>
  <em:spatial>
    <em:country>TGN-ID: 7012149 (United States)</em:country>
    <em:city>TGN-ID: 7014273 (Portland)</em:city>
    <em:region>TGN-ID: 7007708 (Oregon)</em:region>
  </em:spatial>
  <em:epidemiological>
    <em:disease>ICD-10-B03 (Smallpox)</em:disease>
    <em:hostSp>Human</em:hostSp>
    <em:pathoSp>Variola virus</em:pathoSp>
    <em:pathoGroup>Poxviridae</em:pathoGroup>
  </em:epidemiological>
  <em:geographic>Transport network</em:geographic>
  <em:geographic>Movement</em:geographic>
  <em:socioEconomic>Social contact network</em:socioEconomic>
  <em:socioEconomic>Land use</em:socioEconomic>
  <em:bibReference>
    <em:refCitation>S. Eubank, H. Guclu, V.S. Kumar, M.V. Marathe, A.
Srinivasan, Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in
realistic urban social networks.," Nature, vol. 429, 2004, pp. 180-
4.</em:refCitation>
    <em:refDOI>10.1038/nature02541</em:refDOI>
  </em:bibReference>
</em:em>

```

Figure 28- Metadata annotation of a dataset, identified in the paper by Eubank and coworkers, using the EM metadata schema

Chapter 6

Conclusions

The EM will function as data management platform, integrating data from heterogeneous origins and making it available to users and applications according with defined access policies.

The annotation of the resources managed by the repository with semantic rich metadata is essential for data management and exchange between different components of the EM and with external applications. The development of a specific metadata model, making use of controlled vocabularies, such as thesaurus or ontologies, for metadata annotation is an essential step in that direction.

This work has started by the implementation of a digital repository which is available at <http://epiwork.di.fc.ul.pt/muradora>. For this task, an analysis of available software was done. Fedora Commons was chosen due to its capacities to store and organize data and metadata following a digital object model. Moreover, its independence of the internet frontend provides this software with an interesting adaptability.

Muradora was the software chosen to function as an internet frontend for the repository. It provides important features to support the EM requirements, namely: the possibility to use different metadata schemas; advanced options for role based user access management; a search engine; tools for resource browsing and commenting. Furthermore, it supports Shibboleth technology, which is an interesting feature considering the distributed view of the EM as described in section 1.1.

After setting up the digital repository, the EM metadata model was designed with the objective of providing a standardized framework to annotate epidemiological resources with well structured, standardized metadata. This metadata model was based on the DC metadata standard, including specific metadata elements to better structure and support metadata about epidemiological relevant resources. This is done by including, apart from general metadata elements, specific metadata elements to describe important concepts in

the epidemiological field, such as: disease, host species, geographic, and environmental data. The metadata model also defines controlled vocabularies and encoding schemes to support the standardization and semantic enrichment of as many of these metadata elements as possible.

The necessity to include a wide array of metadata elements to describe resources in the repository brings the problem of resistance to metadata creation, due to the effort needed for this task. Thus, another objective of this work was the development of mechanisms to help the user in the annotation process. The inclusion of help menus is one of these mechanisms, which are useful by stating clearly what information is expected in each metadata field. Another of these mechanisms is the use of controlled selection lists, avoiding the use of free text and thus the use of different terms to describe the same concept or orthographic mistakes. Finally some metadata elements can be automatic filled using data available in the EM databases or from the analysis of the uploaded resource.

For the improvement of the help mechanisms available, in the near future, it is necessary to refine the controlled lists available and support them in controlled vocabularies. Metadata re-use is also a concept that needs to be implemented to support the metadata filling process.

The evaluation of the metadata model is fundamental for its improvement. This has been done throughout its development by testing its functionality in the annotation of available datasets or datasets identified in scientific papers. It is important to keep testing the schema as new resources are made available.

Furthermore, it will be important to have feed-back from collaborators and users. The metadata model has been presented to and discussed by the local XLDB Epiwork group. In the future it will be important to have the metadata model tested and evaluated by other Epiwork partners and by the general public, to whom the EM will be made available soon. The metadata model should also be tested in the context of automatic data search and exchange, namely at the level of webservices provided by the mediator.

Finally, after the implementation and refinement of the EM metadata model, this work should evolve to an active involvement in the development of an epidemiological ontology open to the community, following the successful example of GO (Gene ontology).

References

- [1] K. Wilson and J.S. Brownstein, "Early detection of disease outbreaks using the Internet," *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 180, 2009, pp. 829-31.
- [2] M.M. Wagner, F.C. Tsui, J.U. Espino, V.M. Dato, D.F. Sittig, R.A. Caruana, L.F. McGinnis, D.W. Deerfield, M.J. Druzdzal, and D.B. Fridsma, "The emerging science of very early detection of disease outbreaks," *Journal of public health management and practice : JPHMP*, vol. 7, 2001, pp. 51-9.
- [3] S.P. van Noort, M. Muehlen, H. Rebelo De Andrade, C. Koppeschaar, J.M. Lima Lourenço, and M.G. Gomes, "Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe," *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin*, vol. 12, 2007, pp. E5-6.
- [4] I.H. Friesema, C.E. Koppeschaar, G.A. Donker, F. Dijkstra, S.P. van Noort, R. Smalenburg, W. van der Hoek, and M.A. van der Sande, "Internet-based monitoring of influenza-like illness in the general population: experience of five influenza seasons in The Netherlands.," *Vaccine*, vol. 27, 2009, pp. 6353-7.
- [5] R.L. Marquet, A.I. Bartelds, S.P. van Noort, C.E. Koppeschaar, J. Paget, F.G. Schellevis, and J. van der Zee, "Internet-based monitoring of influenza-like illness (ILI) in the general population of the Netherlands during the 2003-2004 influenza season.," *BMC public health*, vol. 6, 2006, p. 242.
- [6] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, 2009, pp. 1012-4.
- [7] J.S. Brownstein and C.C. Freifeld, "HealthMap: the development of automated real-time internet surveillance for epidemic intelligence," *Eurosurveillance*, vol. 12, 2007.
- [8] "Epiwork - The Project. <http://www.epiwork.eu/the-project/>. Accessed at 20-09-2010.."
- [9] L.F. Lopes, F.A. Silva, C. Francisco, J. Zamite, F. Hugo, S. Carla, and M.J. Silva, "Epidemic Marketplace: An Information Management System for Epidemiological Data," *LNCS 6266*, 2010, pp. 31-44.
- [10] J. Zamite, F.A. Silva, F. Couto, and M.J. Silva, "MEDCollector: Multisource Epidemic Data Collector," *Proceedings of ITBAM'10 - 1st International Conference on Information Technology in Bio- and Medical Informatics - DEXA 2010*, 2010.

- [11] L.F. Lopes, J.M. Zamite, B.C. Tavares, F.M. Couto, F. Silva, and M.J. Silva, "Automated Social Network Epidemic Data Collector.," *INForum informatics symposium*, Lisbon: 2009.
- [12] X. Cheng, C. Dale, and J. Liu, "Statistics and Social Network of YouTube Videos," *2008 16th International Workshop on Quality of Service*, 2008, pp. 229-238.
- [13] "Flickr. <http://www.flickr.com/>. Accessed at 13-07-2010."
- [14] B.C. Vanteru, J.S. Shaik, and M. Yeasin, "Semantically linking and browsing PubMed abstracts with gene ontology," *BMC genomics*, vol. 9 Suppl 1, 2008, p. S10.
- [15] H.S. Bilofsky and C. Burks, "The GenBank genetic sequence data bank," *Nucleic acids research*, vol. 16, 1988, pp. 1861-3.
- [16] P. Flicek, B.L. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Gräf, S. Haider, M. Hammond, K. Howe, A. Jenkinson, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, G. Koscielny, E. Kulesha, D. Lawson, I. Longden, T. Massingham, W. McLaren, K. Megy, B. Overduin, B. Pritchard, D. Rios, M. Ruffier, M. Schuster, G. Slater, D. Smedley, G. Spudich, Y.A. Tang, S. Trevanion, A. Vilella, J. Vogel, S. White, S.P. Wilder, A. Zadissa, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernández-Suarez, J. Herrero, T.J. Hubbard, A. Parker, G. Proctor, J. Smith, and S.M. Searle, "Ensembl's 10th year.," *Nucleic acids research*, vol. 38, 2010, pp. D557-62.
- [17] H. Berman, K. Henrick, H. Nakamura, and J.L. Markley, "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data," *Nucleic acids research*, vol. 35, 2007, pp. D301-3.
- [18] The UniProt Consortium, "The Universal Protein Resource (UniProt) in 2010," *Nucleic acids research*, vol. 38, 2010, pp. D142-8.
- [19] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic acids research*, vol. 36, 2008, pp. D344-50.
- [20] A.J. Cuticchia, P. Cooley, R.D. Hall, and Y. Qin, "NIDDK data repository: a central collection of clinical trial data," *BMC medical informatics and decision making*, vol. 6, 2006, p. 19.
- [21] D.C. Hay, *Data model patterns: A metadata map*, San Francisco: Morgan Kaufmann, 2006.
- [22] I. Robu and B. Thirion, "Why and How to Use the Dublin Core Metadata for Health Resources on the Internet: an Introduction," *9th European Conference of Medical and Health Libraries*, Santander: 2004.

- [23] R. Roszkiewicz, "Metadata in Context," *Seybold Report*, vol. 4, 2004.
- [24] V. Brillhante, A. Ferreira, J. Marinho, and P. JS, "Information Integration through Ontology and Metadata for Sustainability Analysis," *Summit on Environmental Modelling and Software Third Biennial*, 2006.
- [25] L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neumann, and S. Stephens, "The semantic web in action," *Scientific American Magazine*.
- [26] M. Rosemann and M.Z. Muehlen, "Evaluation of Workflow Management Systems - A Meta Model Approach," *Australasian Journal of Information Systems*, vol. 6, 1998.
- [27] M. Nagamori and S. Sugimoto, "A Metadata Schema Registry as a Tool to Enhance Metadata Interoperability," *TCDL Bulletin*, vol. 3, 2006.
- [28] Heal, "Heal Metadata Elements Description, version 1.6," 2005, pp. 1-20.
- [29] IMS Global Learning Consortium, "IMS Application Profile Guidelines overview," 2005.
- [30] ISO/IEC, "International Standard: Information technology - Metadata," 2004.
- [31] "METeOR home - <http://meteor.aihw.gov.au/content/index.phtml/itemId/181162>. Accessed at 26-07-2010."
- [32] "United States Health Information Knowledgebase - (USHIK) - <http://ushik.ahrq.gov>. Accessed at 26-07-2010."
- [33] "Cancer Data Standards Registry and Repository (caDSR) - <https://cabig.nci.nih.gov/concepts/caDSR>. Accessed at 26-07-2010."
- [34] D. Powell, M. Nilsson, A. Naeve, P. Johnston, and T. Baker, "DCMI Abstract Model," 2007.
- [35] ISO, *ISO 15836:2009 - Information and documentation - The Dublin Core metadata element set*, Intec, 2009.
- [36] DCMI, "Dublin Core Metadata Element Set, Version 1.1: Reference Description," 2008.
- [37] DCMI Usage Board, "DCMI Metadata Terms," 2008.
- [38] R. Gartner, "MODS: Metadata Object Description Schema," *Pearson New Media Librarian Oxford University Library Services*, vol. 120, 2003, pp. 1-13.
- [39] H. Wagner, "The Dublin Core Metadata Registry. Tools Development With Registry Web Services: An Overview," 2005.

- [40] ISO, *Data elements and interchange formats - Information interchange - Representation of dates and times*, Intec, 2004.
- [41] M. Wolf and C. Wicksteed, "W3C Date and Time Formats," 1998.
- [42] A. Witt, F. Sasaki, E. Teich, N. Calzolari, and P. Wittenburg, "Uses and usage of language resource-related standards," *Proceedings of the LREC 2008 Workshop*, 2008.
- [43] E. Svenonius, "Design of Controlled Vocabularies," *Encyclopedia of Library and Information Science*, New York: Marcel Dekker Inc., 2003.
- [44] T.R. Gruber, *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*, IN FORMAL ONTOLOGY IN CONCEPTUAL ANALYSIS AND KNOWLEDGE REPRESENTATION, KLUWER ACADEMIC PUBLISHERS, IN PRESS. SUBSTANTIAL REVISION OF PAPER PRESENTED AT THE INTERNATIONAL WORKSHOP ON FORMAL ONTOLOGY, 1993.
- [45] O. Bodenreider and R. Stevens, "Bio-ontologies: current trends and future directions.," *Briefings in bioinformatics*, vol. 7, 2006, pp. 256-74.
- [46] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic acids research*, vol. 32, 2004, pp. D267-70.
- [47] "The Open Biological and Biomedical Ontologies - <http://www.obofoundry.org>. Accessed at 7-09-2010."
- [48] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. Sansone, R.H. Scheuermann, N. Shah, P.L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature biotechnology*, vol. 25, 2007, pp. 1251-5.
- [49] L.G. Cowell and B. Smith, "Infectious Disease Ontology," *Infectious disease informatics*, V. Sintchenko, New York: Springer, 2010, pp. 373-395.
- [50] "Yahoo! GeoPlanet™ - YDN - <http://developer.yahoo.com/geo/geoplanet>. Accessed at 26-07-2010."
- [51] M. Wick and T. Becker, "Enhancing RSS feeds with extracted geospatial information for further processing and visualization in The Geospatial Web," *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society* ., A. Scharl and K. Tochtermann, London: Springer-Verlag, 2007, pp. 105-116.
- [52] P. Harpring, "Proper words in proper places: The Thesaurus of Geographic Names," *MDA Information*, vol. 2, 1997, pp. 5-12.
- [53] "European Commission - Inspire - <http://inspire.jrc.ec.europa.eu/>. Accessed at 26-7-2010."

- [54] A.S. Lagoze, C., H. V. Sompel, P. Johnston, M. Nelson, R. Sanderson, "ORE User Guide - Primer," *Open Archives Initiative*, 2008.
- [55] C. Lagoze, S. Payette, E. Shin, and C. Wilper, "Fedora: an architecture for complex objects and their relationships," *International Journal on Digital Libraries*, vol. 6, 2006.
- [56] "National Information Center on Health Services Research and Health Care Technology (NICHSR) - HSR Information Central - <http://www.nlm.nih.gov/hsrinfo/datasites.html>. Accessed at 26-07-2010."
- [57] "CDC Wonder - <http://wonder.cdc.gov/>. Accessed at 26-07-2010."
- [58] D. Revere, P. Bugni, and S. Fuller, "A Public Health Knowledge Management Repository that Includes Grey Literature," *Publishing Research Quarterly*, vol. 23, 2007, pp. 65-70.
- [59] "Clinical Data Repository/Health Data Repository (CHDR) - http://www1.va.gov/VADODHEALTHITSHARING/Clinical_Data_Repository_Health_Data_Repository_CHDR.asp. Accessed at 7-09-2010."
- [60] "Propel Population Health Data Repository - <http://nesstar.uwaterloo.ca/webview/index.jsp>. Accessed at 7-09-2010."
- [61] "ACGT: ACGT - <http://www.eu-acgt.org/>. Accessed at 7-9-2010."
- [62] "caBIG® - <http://cabig.cancer.gov/>. Accessed at 7-09-2010."
- [63] R. Crow, "Open Society Institute - A Guide to Institutional Repository Software," 2004, p. 28.
- [64] M. Branschovsky, R. Lubas, M. Smith, and S. Williams, "Evolving Metadata Needs for an Institutional Repository: MIT's DSpace," *International Conference on Dublin Core and Metadata Applications*, Seattle: 2003, pp. pp. 237-238.
- [65] EPrints, "EPrints version 3 Repository Walkthrough."
- [66] C. Nguyen and J. Dalziel, "Muradora: A Turnkey Fedora GUI Supporting Heterogeneous Metadata, Federated Identity, And Flexible Access Control," *Third International Conference on Open Repositories 2008*, 2008.
- [67] J.C. Ramalho, M. Ferreira, L. Faria, R. Castro, F. Barbedo, and L. Corujo, "RODA and CRiB a service-oriented digital repository," *International Conference on Preservation of Digital Objects*, London: 2008.
- [68] C. Kortekaas, "Don't keep it under your hat: A Fez design description and case study," *Fedora Users Conference 2006*, Charlottesville, Virginia : 2006, pp. 1-21.
- [69] R. Scherle and M. Ozakca, "Developing an Ingest Service for Fedora," *Fedora User Group Meeting at Open Repositories Conference.*, San Antonio: 2007.

- [70] M.A. Leggott, "Islandora: a Drupal/Fedora Repository System," *4th International Conference on Open Repositories*, Georgia: Georgia Institute of Technology, 2009.
- [71] K. Mehlhorn, "EResearch: A Max Planck Perspective," *International Conference on Dublin Core and Metadata Applications.*, Berlin: 2008.
- [72] S. Tuttle, A. Ehlenberger, R. Gorthi, J. Leiserson, R. Macbeth, N. Owen, S. Ranahandola, M. Storrs, and C. Yang, *IBM Redbooks | Understanding LDAP - Design and Implementation*, IBM, 2004.
- [73] OASIS, "eXtensible Access Control Markup Language (XACML) Version 1.0," 2003.
- [74] D. Vohra, "XML.com: Under the Hood: Oracle Berkeley DB XML."
- [75] Y. Seeley, "Apache Solr," *ApacheCon Europe 2006*, Dublin: 2006.
- [76] "GÉANT - <http://www.geant.net/pages/home.aspx>. Accessed at 26-07-2010."
- [77] K. Coyle and T. Baker, "Guidelines for Dublin Core Application Profiles," 2009.
- [78] WHO, "ICD-10. International Statistical Classification of Diseases and Related Health Problems. 10th Revision," 2007.
- [79] "namespaces in xml 1.0 (third edition) - <http://www.w3.org/TR/REC-xml-names/#ns-qualnames>. Accessed at 26-08-2010."
- [80] DCMI-Libraries Working Group, "DC-Library Application Profile (DC-Lib)," 2004.
- [81] N. Ireson, M. Larson, J. Oomen, V. Rondeboom, and H. Smulders, "Metadata schema and mapping – evaluation and revision. Deliverable for project under FP6-2005-IST-5 cal," 2008.
- [82] J.M. Cohen, K.C. Ernst, K.A. Lindblade, J.M. Vulule, C.C. John, and M.L. Wilson, "Topography-derived wetness indices are associated with household-level malaria risk in two communities in the western Kenyan highlands," *Malaria journal*, vol. 7, 2008, p. 40.
- [83] I.J. East, S. Hamilton, and G. Garner, "Identifying areas of Australia at risk of H5N1 avian influenza infection from exposure to migratory birds: a spatial analysis," *Geospatial health*, vol. 2, 2008, pp. 203-13.
- [84] S. Eubank, H. Guclu, V.S. Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in realistic urban social networks.," *Nature*, vol. 429, 2004, pp. 180-4.

Appendix

A.1- EM Property elements based on DCTERMS

From the properties made available by the DCMI, defined in DC-TERMS, a batch that better served the purposes of the Epidemic Marketplace was selected. These terms are individually described in this section.

<i>Name of Term</i>	<i>Abstract</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/abstract
EM qualified name	em:abstract
DC Term URI	http://purl.org/dc/terms/abstract
DC qualified name	dc:abstract
Label	Abstract
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	A summary of the resource
EM Definition	
DC Comments	
EM Comments	
Refines	
Refined By	
Has Encoding Scheme	
Obligation	0
Occurrence	0 or 1

<i>Name of Term</i>	<i>BibliographicCitation</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/bibliographicCitation
EM qualified name	em:bibliographicCitation
DC Term URI	http://purl.org/dc/terms/bibliographicCitation
DC qualified name	dc:bibliographicCitation
Label	Bibliographic Citation
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	A bibliographic reference for the resource.
EM Definition	
DC Comments	Recommended practice is to include sufficient bibliographic detail to identify the resource as unambiguously as possible, whether or not the citation is in a standard form. A draft version of "Guidelines for encoding bibliographic citations in DC metadata" can be found at http://epub.mimas.ac.uk/DC/dc-citation-guidelines/ .
EM Comments	In the future consider a method for the automatic formatting of the

Refines	citation according to a specific format.
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Date</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/date
EM qualified name	em:date
DC Term URI	http://purl.org/dc/terms/date
DC Qualified name	dc:date
Label	Date
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	A date associated with an event in the life cycle of the resource.
EM Definition	Should be the date that the dataset was created or last modified to assume the configuration in which it is submitted to the repository.
DC Comments	Typically, date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and follows the YYYY-MM-DD format.
EM Comments	
Refines	
Refined By	
Has Encoding Scheme	
Obligation	R
Occurrence	1

<i>Name of Term</i>	<i>DateSubmitted</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/dateSubmitted
EM qualified name	em:dateSubmitted
DC Term URI	http://purl.org/dc/terms/dateSubmitted
DC Qualified name	dc:dateSubmitted
Label	Date Submitted
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	Date of submission of the resource (e.g. thesis, articles, etc.).
EM Definition	
DC Comments	
EM Comments	Should be stored automatically by the program when the file is submitted and annotated, without need from user input.
Refines	

Refined By	
Has Encoding Scheme	ISO 8601 - http://purl.org/dc/terms/ISO8601
Obligation	M
Occurrence	0 or 1

<i>Name of Term</i>	<i>Description</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/description
EM qualified name	em:description
DC Term URI	http://purl.org/dc/terms/description
DC qualified name	dc:description
Label	Description
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	An account of the content of the resource.
EM Definition	
DC Comments	Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
EM Comments	
Type of term	Property
Refines	
Refined By	
Has Encoding Scheme	
Obligation	O
Occurrence	0 or 1

<i>Name of Term</i>	<i>Format</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/format
EM qualified name	em:format
DC Term URI	http://purl.org/dc/terms/format
DC Qualified name	dc:format
Label	Format
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	The physical or digital manifestation of the resource.
EM Definition	
DC Comments	Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats).

EM Comments	Should be filled automatically
Refines	
Refined By	
Has Encoding Scheme	IMT - http://purl.org/dc/terms/IMT The Internet media type of the resource: http://www.isi.edu/in-notes/iana/assignments/media-types/media-types
Obligation	M
Occurence	0 or 1

<i>Name of Term</i>	<i>Identifier</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/identifier
EM qualified name	em:identifier
DC Term URI	http://purl.org/dc/terms/identifier
DC Qualified name	dc:identifier
Label	Identifier
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	
EM Definition	
DC Comments	
EM Comments	
Refines	
Refined By	
Has Encoding Scheme	
Obligation	O
Occurence	0 or 1

<i>Name of Term</i>	<i>Language</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/language
EM qualified name	em:language
DC Term URI	http://purl.org/dc/terms/language
DC Qualified name	dc:language
Label	Language
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	A language of the intellectual content of the resource.
EM Definition	
DC Comments	Recommended best practice is to use RFC 3066 [RFC3066], which, in conjunction with ISO 639 [ISO639], defines two- and three-letter primary language tags with optional subtags. Example: en for English, pt for Portuguese
EM Comments	Use the most recent version of the encoding scheme for tags: BCP47. The language name is also included while a system to decode the tag for human users is not implemented.
Refines	
Refined By	

Has Encoding Scheme	BCP47
Obligation	R
Occurrence	0 or more

<i>Name of Term</i>	<i>Publisher</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/publisher
EM qualified name	em:publisher
DC Term URI	http://purl.org/dc/terms/publisher
DC Qualified name	dc:publisher
Label	Publisher
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	An entity responsible for making the resource available.
EM Definition	In the Epidemic marketplace it is user that submits the resource to the repository.
DC Comments	Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
EM Comments	The user may be a person or an organization. Should be filled automatically.
Refines	
Refined By	http://epiwork.di.fc.ul.pt/em-terms/pubname http://epiwork.di.fc.ul.pt/em-terms/puborg http://epiwork.di.fc.ul.pt/em-terms/puburl
Has Encoding Scheme	
Obligation	R
Occurrence	0 or 1

<i>Name of Term</i>	<i>Rights</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/rights
EM qualified name	em:rights
DC Term URI	http://purl.org/dc/terms/rights
DC Qualified name	dc:rights
Label	Rights
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	Information about rights held in and over the resource.
EM Definition	
DC Comments	Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource.
EM Comments	
Refines	

Refined By	http://epiwork.di.fc.ul.pt/em-terms/rightsHolder http://epiwork.di.fc.ul.pt/em-terms/copyright http://epiwork.di.fc.ul.pt/em-terms/disclaimer
Has Encoding Scheme	
Obligation	RA
Occurence	0 or 1

<i>Name of Term</i>	<i>RightsHolder</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/rightsHolder
EM qualified name	em:rightsHolder
DC Term URI	http://purl.org/dc/terms/rightsHolder
DC Qualified name	dc:rightsHolder
Label	Rights Holder
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	A person or organization owning or managing rights over the resource.
EM Definition	
DC Comments	
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/rights
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurence	0 or more

<i>Name of Term</i>	<i>Source</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/source
EM qualified name	em:source
DC Term URI	http://purl.org/dc/terms/source
DC Qualified name	dc:source
Label	Source
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	A Reference to a resource from which the present resource is derived.
EM Definition	
DC Comments	
EM Comments	Description of the source structured in the sub-elements source name, description and URL
Refines	
Refined By	http://epiwork.di.fc.ul.pt/em-terms/srcName http://epiwork.di.fc.ul.pt/em-terms/srcDescription http://epiwork.di.fc.ul.pt/em-terms/srcURL

Has Encoding Scheme Obligation Occurrence	RA 0 or more
---	------------------------

<i>Name of Term</i>	<i>Spatial</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/spatial
EM qualified name	em:spatial
DC Term URI	http://purl.org/dc/terms/spatial
DC Qualified name	dc:spatial
Label	Spatial
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	Spatial coverage of the resource.
EM Definition	
DC Comments	
EM Comments	
Refines	
Refined By	http://epiwork.di.fc.ul.pt/em-terms/city http://epiwork.di.fc.ul.pt/em-terms/country http://epiwork.di.fc.ul.pt/em-terms/region
Has Encoding Scheme	Yahoo! GeoPlanet DCMI Point - http://purl.org/dc/terms/Point ISO 3166 - http://purl.org/dc/terms/ISO3166 DCMI Box - http://purl.org/dc/terms/Box TGN - http://purl.org/dc/terms/TGN Geonames
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Subject</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/subject
EM qualified name	em: subject
DC Term URI	http://purl.org/dc/terms/subject
DC Qualified name	dc:subject
Label	Subject
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	The topic of the content of the resource.
EM Definition	Use keywords
DC Comments	Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.
EM Comments	
Refines	
Refined By	

Has Encoding Scheme Obligation Occurence	R 0 or more
--	-----------------------

<i>Name of Term</i>	<i>Temporal</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/temporal
EM qualified name	em:temporal
DC Term URI	http://purl.org/dc/terms/temporal
DC Qualified name	dc:temporal
Label	Temporal
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	Temporal characteristics of the resource.
EM Definition	
DC Comments	
EM Comments	
Refines	
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurence	0 or more

<i>Name of Term</i>	<i>Title</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/title
EM qualified name	em:title
DC Term URI	http://purl.org/dc/terms/title
DC Qualified name	dc:title
Label	Title
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	A name given to the resource
EM Definition	-
DC Comments	Typically, a title will be a name by which the resource is formally known.
EM Comments	If no title is given a name should be assigned automatically.
Refines	
Refined By	
Has Encoding Scheme	
Obligation	M
Occurence	1

<i>Name of Term</i>	<i>Type</i>
EM Term URI	http://epiwork.di.fc.ul.pt/em-terms/type
EM qualified name	em:type
DC Term URI	http://purl.org/dc/terms/type
DC Qualified name	dc:type
Label	Type
Defined By	http://dublincore.org/documents/dcmi-terms
DC Definition	The nature or genre of the content of the resource.
EM Definition	
DC Comments	Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of DCMI Types). To describe the physical or digital manifestation of the resource, use the Format element.
EM Comments	Use a controlled list
Refines	
Refined By	
Has Encoding Scheme	EM-Type
Obligation	M
Occurrence	1

A.2- EM metadata elements defined locally

In this section several metadata property elements are proposed to be used specifically in the Epidemic Marketplace (EM). These terms should describe epidemiological and related data.

<i>Name of Term</i>	<i>Author</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/author
Qualified name	em:author
Label	Author
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Set of metadata elements to describe the author(s).
EM Comments	
Refines	
Refined By	http://epiwork.di.fc.ul.pt/em-terms/authname http://epiwork.di.fc.ul.pt/em-terms/authororg http://epiwork.di.fc.ul.pt/em-terms/authorurl
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>AuthName</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/authName
Qualified name	em:authName
Label	Author Name.
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The name of the author.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/author
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or 1

<i>Name of Term</i>	<i>AuthOrg</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/authOrg
Qualified name	em:authOrg
Label	Author Affiliation.
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The organization for which the author works for.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/author
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or 1

<i>Name of Term</i>	<i>AuthorURL</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/authorURL
Qualified name	em:authorURL
Label	Author Homepage
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Author homepage
EM Comments	Use URL
Refines	http://epiwork.di.fc.ul.pt/em-terms/author
Refined By	
Has Encoding	RFC 3986
Scheme	URI - http://purl.org/dc/terms/URI
Obligation	RA
Occurence	0 or 1

<i>Name of Term</i>	<i>Citation</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/citation
Qualified name	em:citation
Label	Citation
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Bibliographic reference of the resource. This applies for published resources, such as books, journals and technical reports.
EM Comments	Unlike bibliographicCitation, this is resource citation The use of a specific format, such as bibTEX, or other widely used is recommended.
Refines	http://epiwork.di.fc.ul.pt/em-terms/generalDescription
Refined By	
Has Encoding	
Scheme	
Obligation	RA
Occurence	0 or 1

<i>Name of Term</i>	<i>City</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/city
Qualified name	em:city
Label	City
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A city.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/spatial
Refined By	
Has Encoding	Yahoo! GeoPlanets
Scheme	
Obligation	RA
Occurence	0 or more

<i>Name of Term</i>	<i>Copyright</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/copyright
Qualified name	em:copyright
Label	Copyright
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The copyright of the resource.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/rights
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or 1

<i>Name of Term</i>	<i>Country</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/country
Qualified name	em:country
Label	Country
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A country.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/spatial
Refined By	
Has Encoding Scheme	Yahoo! GeoPlanets
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Demographic</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/demographic
Qualified name	em:demographic
Label	Demographic
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Describes the type of demographic elements contained in the resource. For example, population size, distribution, birth rates, etc.
EM Comments	A controlled list of demographic elements should be defined.
Refines	
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Disclaimer</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/disclaimer
Qualified name	em:disclaimer
Label	Disclaimer
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A disclaimer for the resource.

EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/rights
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or 1

<i>Name of Term</i>	<i>DiagnosticMethod</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/diagnosticMethod
Qualified name	em:diagnosticMethod
Label	Diagnostic Method
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A diagnostic test used to obtain data in the resource or which protocol is defined in the resource.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Disease</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/disease
Qualified name	em:disease
Label	Disease
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A unequivocal name for a disease (or more) about which data is contained in the resource
EM Comments	The ICD-10 disease coding or the UMLS (which includes ICD) should be used for the standardization the disease names used.
Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	ICD-10 UMLS
	http://www.infectiousdiseaseontology.org/IDO.html http://code.google.com/p/infectious-disease-ontology/ http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>DOI</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/DOI
Qualified name	em:DOI
Label	DOI
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The DOI of the resource.

EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/generalDescription
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or 1

<i>Name of Term</i>	<i>Drug</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/drug
Qualified name	em:drug
Label	Drug
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Describe chemical compounds used in disease treatment.
EM Comments	A controlled vocabulary should be used. While not found should be user defined.
Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	National drug Code Directory (FDA) http://www.accessdata.fda.gov/scripts/cder/ndc/default.cfm ChEBI
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Environmental</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/environmental
Qualified name	em:environmental
Label	Environmental
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Describe environmental data contained in the resource.
EM Comments	A list of environmental topics should be provided.
Refines	
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Epidemiological</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Qualified name	em:epidemiological
Label	Epidemiological
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A group of metadata elements for the description of epidemiological data contained in the resource.
EM Comments	
Refines	
Refined By	http://epiwork.di.fc.ul.pt/em-terms/diagnosticMethod http://epiwork.di.fc.ul.pt/em-terms/disease

Has Encoding Scheme	http://epiwork.di.fc.ul.pt/em-terms/drug http://epiwork.di.fc.ul.pt/em-terms/hostSp http://epiwork.di.fc.ul.pt/em-terms/hostgroup http://epiwork.di.fc.ul.pt/em-terms/pathSp http://epiwork.di.fc.ul.pt/em-terms/pathoGroup http://epiwork.di.fc.ul.pt/em-terms/strain http://epiwork.di.fc.ul.pt/em-terms/vaccine http://epiwork.di.fc.ul.pt/em-terms/vector
Obligation	RA
Occurence	0 or more

<i>Name of Term</i>	<i>HostSp</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/hostSp
Qualified name	em:hostSp
Label	Host Species
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The species of the organism that is the disease host.
EM Comments	Define controlled vocabulary. Taxonomy. A scientific classification should be used- species name. Ex.: Homo sapiens sapiens
Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	http://www.treebase.org/treebase-web/home.html;jsessionid=9F231B51776C8038FC805F71BA3E1D23
Obligation	RA
Occurence	0 or more

<i>Name of Term</i>	<i>HostGroup</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/hostGroup
Qualified name	em:hostGroup
Label	Host Group
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A dataset may contain data not about a specific species but rather a larger group comprising different species. When the larger group is the target this element should be used. For example, when referring to groups such as rodents or household animals.
EM Comments	When possible a scientific classification should be used.
Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurence	0 or more

<i>Name of Term</i>	<i>GeneralDescription</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/generalDescription
Qualified name	em:generalDescription

Label	General Description
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Group of metadata elements to provide a general description of the resource.
EM Comments	A controlled list needs to be defined
Refines	
Refined By	http://epiwork.di.fc.ul.pt/em-terms/abstract http://epiwork.di.fc.ul.pt/em-terms/description http://epiwork.di.fc.ul.pt/em-terms/citation http://epiwork.di.fc.ul.pt/em-terms/DOI http://epiwork.di.fc.ul.pt/em-terms/format http://epiwork.di.fc.ul.pt/em-terms/identifier http://epiwork.di.fc.ul.pt/em-terms/ISBN http://epiwork.di.fc.ul.pt/em-terms/ISSN http://epiwork.di.fc.ul.pt/em-terms/language http://epiwork.di.fc.ul.pt/em-terms/pubmedID http://epiwork.di.fc.ul.pt/em-terms/type http://epiwork.di.fc.ul.pt/em-terms/typeOfDoc http://epiwork.di.fc.ul.pt/em-terms/typeOfWR http://epiwork.di.fc.ul.pt/em-terms/URL http://epiwork.di.fc.ul.pt/em-terms/venue http://epiwork.di.fc.ul.pt/em-terms/version
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Geographic</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/geographic
Qualified name	em:geographic
Label	Geographic
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Describe geographic data contained in the resource.
EM Comments	A list of environmental topics should be provided.
Refines	
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>ISBN</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/ISBN
Qualified name	em:ISBN
Label	ISBN
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The ISBN number of the resource.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/generalDescription

Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>ISSN</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/ISSN
Qualified name	em:ISSN
Label	ISSN
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The ISSN number of the resource.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/generalDescription
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Organization</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/organization
Qualified name	em:organization
Label	Organization
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Set of element to describe the organization responsible for the creation of a resource.
EM Comments	To be used instead of author, when the author is presented as an organization instead of a person or group of people.
Refines	
Refined By	http://epiwork.di.fc.ul.pt/em-terms/orgname http://epiwork.di.fc.ul.pt/em-terms/orgURL
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>OrgName</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/orgname
Qualified name	em:orgName
Label	Organization Name
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The name of the organization responsible for the creation of the dataset.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/organization
Refined By	
Has Encoding Scheme	

Obligation Occurrence	RA 0 or more
-----------------------	------------------------

<i>Name of Term</i>	<i>OrgURL</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/orgURL
Qualified name	em:orgURL
Label	Organization web address
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The web address of organization responsible for the creation of the dataset.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/organization
Refined By	
Has Encoding Scheme	RFC 3986
Obligation Occurrence	RA 0 or more

<i>Name of Term</i>	<i>pathoGroup</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/pathoGroup
Qualified name	em:pathoGroup
Label	Pathogen Group
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A dataset may contain data not about a specific species but rather a larger group comprising different species. When the larger group is the target this element should be used. For example when referring to parasites of the Apicomplexa Phylum.
EM Comments	When possible a scientific classification should be used.
Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	
Obligation Occurrence	RA 0 or more

<i>Name of Term</i>	<i>PathoSp</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/pathoSp
Qualified name	em:pathoSp
Label	Pathogen species
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Identifies a pathogenic organism. Ex.: <i>Staphylococcus aureus</i> , <i>Haemophilus influenza</i> , <i>Mycobacterium tuberculosis</i> , <i>Plasmodium falciparum</i>
EM Comments	A scientific classification should be used- species name.
Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	http://www.viprbrc.org/brc/home.do?decorator=vipr http://eupathdb.org/eupathdb/ http://www.issg.org/database/species/ecology.asp?fr=1&si=326&sts

	http://www.dsmz.de/microorganisms/main.php?contentleft%20id=14 http://www.species2000.org/ http://www.itis.gov/ http://doi.namesforlife.com/ http://bac.hs.med.kyoto-u.ac.jp/alphabet-e.html
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>PathoStrain</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/pathoStrain
Qualified name	em:pathoStrain
Label	Pathogen Strain
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Identifies a pathogenic organism. Ex.: <i>Staphylococcus aureus</i> , <i>Haemophilus influenza</i> , <i>Mycobacterium tuberculosis</i> , <i>Plasmodium falciparum</i>
EM Comments	A scientific classification should be used- species name.
Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>PubmedID</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/pubmedID
Qualified name	em:pubmedID
Label	Pubmed ID
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The ID in the Pubmed repository
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/generalDescription
Refined By	
Has Encoding Scheme	
Obligation	O
Occurrence	0 or 1

<i>Name of Term</i>	<i>PubName</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/pubName
Qualified name	em:pubName
Label	Publisher Name
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Name of the user that submits the resource to the EM.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/publisher

Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>PubOrg</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/pubOrg
Qualified name	em:pubOrg
Label	Publisher Affiliation
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Organization for which the publisher works for or if the user who submits is an organization.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/publisher
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>PubURL</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/pubURL
Qualified name	em:pubURL
Label	Publisher Homepage
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Homepage of the publisher.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/publisher
Refined By	
Has Encoding Scheme	RFC 3986
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Region</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/region
Qualified name	em:region
Label	Region
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A region.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/spatial
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>RefCitation</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/refCitation
Qualified name	em: refCitation
Label	Reference Citation
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The citation of a reference for the resource.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/bibliographicCitation
Refined By	
Has Encoding Scheme	
Obligation	O
Occurrence	0 or 1

<i>Name of Term</i>	<i>RefDOI</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/refDOI
Qualified name	em: refDOI
Label	Reference DOI
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The DOI of a reference for the resource.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/bibliographicCitation
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or 1

<i>Name of Term</i>	<i>RefPubmedID</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/refPubmedID
Qualified name	em:refPubmedID
Label	Reference Pubmed ID
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The citation of a reference for the resource.
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/bibliographicCitation
Refined By	
Has Encoding Scheme	
Obligation	O
Occurrence	0 or 1

<i>Name of Term</i>	<i>SocioEconomic</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/socioEconomic
Qualified name	em:socioEconomic
Label	Socio-Economic
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Describe social or/and economical data contained in the resource.
EM Comments	A list of topics should be provided.

Refines	
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>SrcDescription</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/srcDescription
Qualified name	em:srcDescription
Label	Source Description
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A description of the source of the resource
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/source
Refined By	
Has Encoding Scheme	
Obligation	O
Occurrence	0 or 1

<i>Name of Term</i>	<i>srcName</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/srcName
Qualified name	em:srcName
Label	Source Name
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The name of the source
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/source
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or 1

<i>Name of Term</i>	<i>SrcURL</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/srcURL
Qualified name	em:srcURL
Label	Source URL
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	The URL of the source
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/source
Refined By	
Has Encoding Scheme	RFC 3986 URI - http://purl.org/dc/terms/URI
Obligation	RA
Occurrence	0 or 1

<i>Name of Term</i>	<i>TempFrom</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/tempFom
Qualified name	em: tempFrom
Label	From (date)
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A date relative to the start of a period of time.
EM Comments	Used for the description of time periods related to the temporal coverage. Requires the filling of tempTo.
Refines	http://epiwork.di.fc.ul.pt/em-terms/temporal
Refined By	
Has Encoding Scheme	W3C-DTF - http://purl.org/dc/terms/W3CDTF
Obligation	RA
Occurence	0 or more

<i>Name of Term</i>	<i>TempTo</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/tempTo
Qualified name	em: tempTo
Label	To (date)
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	A date relative to the end of a period of time.
EM Comments	Used for the description of time periods related to the temporal coverage. Requires the filling of tempFrom.
Refines	http://epiwork.di.fc.ul.pt/em-terms/temporal
Refined By	
Has Encoding Scheme	W3C-DTF - http://purl.org/dc/terms/W3CDTF
Obligation	RA
Occurence	0 or more

<i>Name of Term</i>	<i>TypeOfDoc</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/typeofdoc
Qualified name	em: typeOfDoc
Label	Type of Document
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Type of document.
EM Comments	A list of types of documents should be provided.
Refines	
Refined By	
Has Encoding Scheme	EM-Doc-Type
Obligation	RA
Occurence	0 or more

<i>Name of Term</i>	<i>TypeOfWR</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/typeofwr
Qualified name	em: typeOfWR
Label	Type of Web Resource
Defined By	http://epiwork.di.fc.ul.pt/em-terms

EM Definition	Type of Web Resource
EM Comments	A list of types of web resources should be provided.
Refines	
Refined By	
Has Encoding Scheme	EM-WR-Type
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>URL</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/URL
Qualified name	em:URL
Label	URL
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	URL of the resource.
EM Comments	This is to be applied to web resources that have a specific URL. Different from source URL.
Refines	http://epiwork.di.fc.ul.pt/em-terms/generalDescription
Refined By	
Has Encoding Scheme	RFC 3986
Scheme	URI - http://purl.org/dc/terms/URI
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Vaccine</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/vaccine
Qualified name	em:vaccine
Label	Vaccine
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Describe a vaccine used in the study about which the dataset refers. Ex.: data about vaccination tests
EM Comments	Vaccine compound or name by which it is known.
Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	Vaccine ontology
Scheme	http://www.violinet.org/vaccineontology/
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Vector</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/vector
Qualified name	em:vector
Label	Vector
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Identifies the organism that is the vector of a specific disease covered in the dataset.
EM Comments	A scientific classification should be used when possible. Ex.: a species name (ex.: <i>Anopheles gambiae</i>) or a group (anopheline mosquitoes)

Refines	http://epiwork.di.fc.ul.pt/em-terms/epidemiological
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Venue</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/venue
Qualified name	em:venue
Label	Venue
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Specific location of an event, such as an address
EM Comments	
Refines	http://epiwork.di.fc.ul.pt/em-terms/generalDescription
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more

<i>Name of Term</i>	<i>Version</i>
Term URI	http://epiwork.di.fc.ul.pt/em-terms/version
Qualified name	em:version
Label	Version
Defined By	http://epiwork.di.fc.ul.pt/em-terms
EM Definition	Defines the resource version.
EM Comments	The syntax to be used should conform to the versioning system used by the author.
Refines	http://epiwork.di.fc.ul.pt/em-terms/generalDescription
Refined By	
Has Encoding Scheme	
Obligation	RA
Occurrence	0 or more