From Ontology Selection and Semantic Web to an Integrated Information System for Food-borne Diseases and Food Safety

Xianghe Yan[1,*], Yun Peng[2], Jianghong Meng[3], Juliana Ruzante[3], Pina M. Fratamico[1], Lihan Huang[1], Vijay Juneja[1,] David S. Needleman[1]

[1]U.S. Department of Agriculture, Agricultural Research Service, Eastern Regional Research Center, Wyndmoor, PA 19038,

[2]Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore, Maryland 21250,

[3]Joint Institute for Food Safety and Applied Nutrition (JIFSAN), University of Maryland, College Park, MD 20742

*Corresponding author: 215-233-6732 (Phone), 215-233-6581 (Fax), E-mail: Xianghe.yan@ars.usda.gov

Abstract Several factors have hindered effective use of information and resources related to food safety due to inconsistency among semantically heterogeneous data resources, lack of knowledge on profiling of food-borne pathogens, and knowledge gaps among research communities, government risk assessors/managers, and end users of the information. This paper discusses technical aspects in the establishment of a comprehensive food safety information system consisting of the following steps: a) computational collection and compiling publicly available information, including published pathogen genomic, proteomic, and metabolomic data; b) development of ontology libraries on food-borne pathogens and design automatic algorithms with formal inference and fuzzy and probabilistic reasoning to address the consistency and accuracy of distributed information resources (e.g., PulseNet, FoodNet, OutbreakNet, PubMed, NCBI, EMBL, and other online genetic databases and information); c) integration of collected pathogen profiling data, Foodrisk.org (http://www.foodrisk.org), PMP, Combase, and other relevant information into a user-friendly, searchable, "homogeneous" information system available to scientists in academia, the food industry, and government agencies; and d) development of a computational model in semantic web for greater adaptability and robustness.

INTRODUCTION

Food-borne illness is an important public health concern in developing, as well as developed countries. Prevention of food-borne illness and outbreaks through effective interventions, availability of early warning systems, and reliable detection methods for food-borne pathogens is a critical issue worldwide. It is estimated that food-borne pathogens cause approximately 76 million cases of gastrointestinal illnesses, 325,000 hospitalizations, and 5,000 deaths in the United States annually [2, 18].

Over the last three decades, remarkable advances in information and communication technologies (ICTs), genomics and other cutting-edge “omics” technologies have dramatically improved our ability to rapidly determine and interpret the mechanisms of survival and pathogenesis of human food-borne pathogens. Data collection, analysis, and the timely dissemination of these data are essential components for the planning, implementation, and evaluation of public health practices. There are over numerous important mechanisms (Uniform Resource Locators [URLs]) for data sharing and accessing of food safety information, related to microbial and chemical contamination, pathogen characteristics and predictive microbiology, public health surveillance, risk assessment and risk analysis, inspection, management and regulation, recalls, violations, prevention and control, and others [26]. However, a centralized information system to handle the data flow from these information resources, to standardize the content of these resources and to integrate this information with data in public repositories is sorely lacking.

The purpose of this paper is to discuss how an integrated information system could be used to integrate data from heterogeneous resources to strengthen food-borne pathogen risk management, surveillance, and prevention systems and to lay the groundwork for a standard interoperable protocol that could serve as a nationwide food-borne pathogen-related warning system.

CHALLENGES

There are many challenges associated with establishing a centralized Food Safety Information Reporting System (FSIRS), including data access issues, standards and data format issues. One of the largest challenges when creating a FSIRS is accurate and reliable prediction of the combined effects of complex multi-factorial factors on the growth and inactivation of food-borne pathogens. The development of predictive models involves conducting extensive scientific experiments to investigate the biological behaviors of microorganisms under a variety of conditions and fitting the experimental data into appropriate primary and secondary models. Data sharing is another large challenge facing the development of FSIRS. Data mining and large scale statistical data analysis is a time-consuming process. The presentation of food-borne pathogen surveillance and prevention systems must be accurate and be ready to detect changes in complex heterogeneous data systems very quickly. This requires advanced algorithms, data structures, and dynamic communication tools (e.g. web) for detection and prediction of transmission patterns of food-borne pathogens. Advances in algorithms, data structures, and artificial intelligence allow for practical applications of data-driven outbreak detection methods, which can handle the complexity of the task at hand by learning from examples in historical data and from real or simulated recorded outbreaks. The Semantic Web first was considered by Tim Berners-Lee, inventor of the WWW, URIs, HTTP, and HTML in early 2001 "-semantics is considered to be the best framework to deal with the heterogeneity, massive scale, and dynamic nature of the resources on the Web" [23]. The semantic web services can search and evoke over thousands of internet URLs quickly to embrace multi-inputs and multi-outputs for accurate and rapid food-borne pathogen emergency decision making. Several information technologies such as the extensible Markup Language (XML), the Web Ontology Language (OWL), and the Resource Description Framework (RDF) could be used to reduce the syntactic diversity and enable the system to manipulate and make inferences about the data, thereby defining meaningful relationships between the items. A deductive query language based on the semantic web ontology, OWL-QL, or other advanced technologies could facilitate user queries. The extended semantic markup language should be considered to represent both logical and probabilistic relations in web resources on a unified semantic basis. The developed reasoning methods are capable of resolving semantic differences between web resources, making probabilistic reasoning over the semantic web possible. Capability of resolving semantic differences is a significant advance in the area of distributed BN where existing methods all rely on exchanging beliefs via shared identical variables. This work also offers a solution to a persistent problem facing current semantic web ontologies, namely, their inability to represent and reason upon uncertain relations and inputs. In particular, treating probabilistically enhanced OWL ontologies as probabilistic resources, concept mapping between ontologies can be accomplished as evidential reasoning using the developed reasoning method. From both surveillance and prevention points of views, some critical challenges need to be addressed:

Heterogeneous Data Representation: There is a tremendous amount of online scientific literature, regulations, and pathogen profiling data. However, there is no standard language to represent semantics and heterogeneities of mined knowledge in the semantically heterogeneous scenario. Although manual translation is possible, it is time consuming and error prone if the size of knowledge is large.

The Correctness and Accuracy of Knowledge Prediction: The key problems associated with the correctness and accuracy of knowledge prediction is short of a unified data annotation ontology standard, corrected applied algorithms, and timely entry of public health data. Therefore, it is important to set up an ontology standard and develop a dynamic user interface by using semantic web technology**.**

Timing of Food Safety Emergency Responses: An important part of food-borne pathogen surveillance and prevention is the timing of food safety emergency responses. In some cases, determining the geographical scope of food safety emergency and/or investigating the course of the food safety "crisis" is difficult. Therefore, it is important to quickly compile multivariate data (ideally, in real time) through neural network analysis to maintain the data correlation ability of prediction hypotheses, regulatory threshold, and the individual cases. Advances in algorithms, data ontology analysis, and neural network allow for practical application of data-driven outbreak investigation methods.

DATA SOURCES

Pathogen profiling of food-borne pathogens: The accurate classification of food-borne pathogens through integration of microbial genomics data with clinical and phenotypic observations is an inportant tool for the detection of patterns of food-borne pathogen transmission and the construction of epidemic trees.

PMP: The Pathogen Modeling Program (PMP), established and maintained at the USDA-ARS-Eastern Regional Research Center (ERRC), is a package of models that can be used to predict the growth, survival, and inactivation of food-borne bacteria, primarily pathogens, under various environmental conditions [17, 22, and 27]. The PMP is periodically updated as new models and/or changes in the model-user interface become available.

Combase: This freely accessible Combined Database of Predictive Microbiology Information (ComBase) database was jointly developed by the USDA-ARS-ERRC, the Institute of Food Research, Norwich, U.K., Food Standards Agency, London, U.K., and the Food Safety Centre, Australia. ComBase [3 and 4] contains data sets submitted by researchers and from publications that describe the rate of growth, survival, and inactivation of bacteria under diverse environments relevant to food processing operations.

PMIP portal: The Predictive Microbiology Information Portal (PMIP) contains three major sections [14], which provide access to predictive models for food-borne pathogens, relevant regulatory policies and guidelines, and microbial data related to pathogenic and spoilage microorganisms in a wide variety of food products.

Foodrisk.org: Foodrisk.org is an online resource managed by the Joint Institute for Food Safety and Applied Nutrition (JIFSAN), a joint Institute between the United States Food and Drug Administration (FDA) and the University of Maryland at College Park. Some of the content of Foodrisk.org is unique to the website and can only be found there.

Other online information resources: CDC’s databases (PulseNet, FoodNet, and OutbreakNet) are already used to detect outbreaks early based on pathogen pulsed-field gel electrophoresis (PFGE) patterns.

ROADMAP TO INTEGRATED INFORMATION SYSTEM OF FOOD-BORNE DISEASES AND FOOD SAFETY

This FSIRS will be based on three different methodologies: Neural Network analysis, Bayesian Network Modeling, and Semantic Web (OWL/RDF) technology. The input to this system will include risk assessment information, document (rule)-based regulation policies, publications on experimental-based pathogen profiling, large scale statistical data sets, and predictive modeling from real-time data and/or simulated outbreak analyses. The overall execution and management of this information system will be divided into three main sub-stages as illustrated in **Fig. 1**.

Annotation Stage (**Fig. 1**): Ontology development is the fundamental part of semantic web. Web search engines and the richest information resources such as Google using keywords, PubMed (indexed for MEDLINE) of NCBI, the European Molecular Biology Laboratory (EMBL), DNA Databank of Japan (DDBJ) through terminologies, and the most usable food safety knowledge via open URLs, such as Combase, foodrisk.org, PulseNet (http://www.cdc.gov/pulsenet/), Food Net (http://www.cdc.gov/FoodNet/), Outbreak Net (http://www.cdc.gov/OutbreakNet/), etc. to automatically retrieve text exemplars for each standardized ontology concept from the web could be used at this stage. Here, ontology is a formal, explicit description of concepts, their properties, and relationships among the terminology of food-borne pathogens. Note that the definition of terminology and standardized ontology will not generated by Data Analysts alone, instead it will be a widely agreed-upon standard determined by experts in different fields. All these documents will then be classified into individual subcategories based on classifications related to food-borne pathogens and food safety.

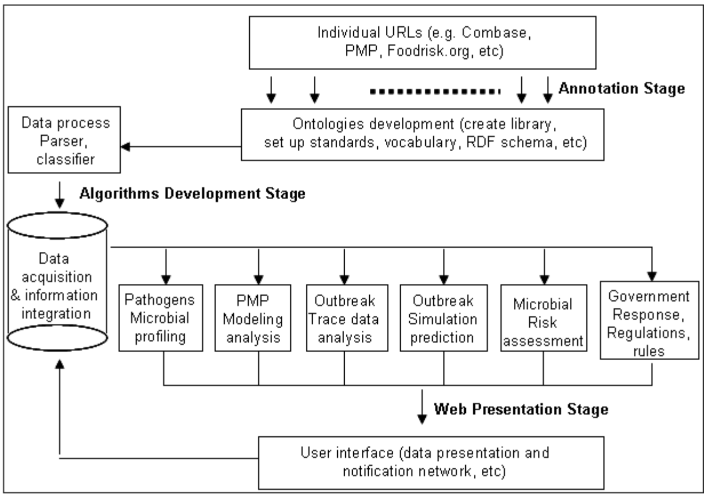
Individual URLs (e.g. Combase, PMP, Foodrisk.org, etc)
Annotation Stage
Data process Parser, classifier
Ontologies development (create library, set up standards, vocabulary, RDF schema, etc)
Algorithms Development Stage
Data acquisition & information integration
Pathogens Microbial profiling
PMP Modeling analysis
Outbreak Trace data analysis
Outbreak Simulation prediction
Microbial Risk assessment
Government Response, Regulations, rules
Web Presentation Stage
User interface (data presentation and notification network, etc)

Figure 1: Semantic Ontology Driven Information System Architecture

Fig. 2 describes the details of the food-borne pathogen ontology categories and subject descriptors. The flow chart of the categories and subject descriptors for food-borne pathogen ontology development in **Fig. 2** not only includes the definition of classes, relationships, and attributes, but also defines a set of subcategories for further classification. To guarantee relevancy, the search will be guided by semantic information given in the ontology where the concept is defined. For example, instead of using "***Escherichia coli* O157:H7**", the term for the concept, as search query, we can form the query by including all terms on the path from the root to this class in the ontology, e.g., **"food-borne pathogen + bacteria +**

Escherichia coli* O157:H7"** for positive exemplars and "**food-borne pathogen + bacteria – *Escherichia coli* O157:H7**" for negative exemplars. Other ontological information can also be included, e.g., "**non-pathogen**" which is another super class of **bacteria**, "**pink**" which is the value for *Escherichia coli* O157:H7 colony color cultured on the agar medium called sorbitol MacConkey agar (SMAC) (a property), and "**toxB**" which is a property inherited from "Escherichia coli* O157:H7"**.

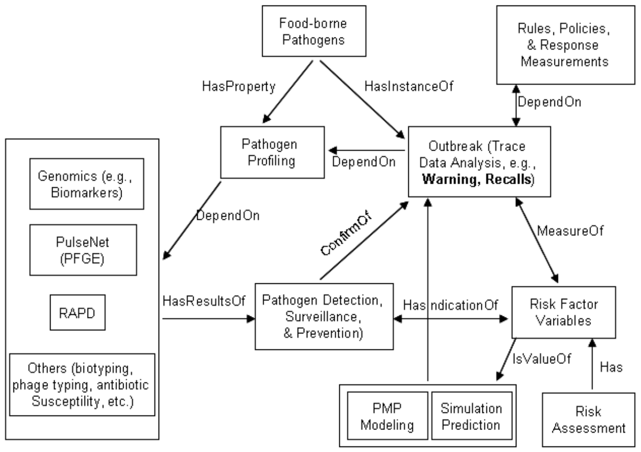
Food-borne Pathogens
Rules, Policies, & Response Measurements
HasProperty
HasInstanceOf
DependOn
Pathogen Profiling
DependOn
Outbreak (Trace Data Analysis, e.g., Warning, Recalls)
Genomics (e.g., Biomarkers)
DependOn
ConfirmOf
MeasureOf
PulseNet (PFGE)
RAPD
HasResultsOf
Pathogen Detection, Surveillance, & Prevention)
HasIndicationOf
Risk Factor Variables
IsValueOf
Has
Others (biotyping, phage typing, antibiotic Susceptility, etc.)
PMP Modeling
Simulation Prediction
Risk Assessment

Figure 2: The Categories and Subject Descriptors for Food-borne Pathogen Ontology Development.

Algorithms Development Stage (Fig. 1): Research in this direction could focus on methods to form search queries that best utilize available semantic information. Questions that will be addressed include, for example, what semantic information should be included; how to order the terms so that their semantic distances to the concept is reflected; and if and how to form alternative queries to represent disjunctive relations and synonyms [9, 10, 15, 19, and 21]. In addition, methods for processing returned pages to filter less relevant ones by, say, various data mining techniques such as clustering will be investigated. Research methods will be empirical. Computer experiments will be conducted to assess if and how much improvement can be gained with a particular technique. To formally categorize and resolve the semantic heterogeneity problems in food safety information resources, an ontology mining framework will first be to extend to discover semantics from both relational data and semi-structured data and represent the mined knowledge with ontologies, which have been used for the formal specification of conceptualization in traditional knowledge engineering and the emerging Semantic Web. The heterogeneities between different data resources will be represented as formal ontological mappings. The mappings could be specified by humans manually or discovered by mapping tools automatically. However, the automatically discovered mappings will have some uncertainty. It should provide a formal representation of uncertain heterogeneities as fuzzy and probabilistic mappings by distinguishing subjective perceptions from objective measurements.

From the distributed computing point of view, engineers may need to design knowledge translation algorithms for distributed data mining (DDM) systems in a client-server model. The clients will be data analysts from a domain (e.g., food processing industries or government agencies) and a DDM server will connect to local data resources. A software engineer will also ensure that the server will include a meta-data (e.g., ontologies and mappings) repository. The principle for this knowledge translation algorithm design is that the only thing transferred from local individual information resources (local miners) to the user (server) side is the translated knowledge through the Ontology & Mapping Repository. Engineers should avoid transferring data back and forth in order to achieve communication efficiency and good scalability. The data mining tasks could run on local resources to achieve locality of the computation. The XML namespace (XML+NS), XML schema, and RDF schema will be developed at this stage (**Fig. 3**).

Web Presentation Stage (Fig. 1): The semantic web (SW) extends the current web by specifying the semantics or meaning of information on the web using markup languages so that it can be understood and processed not only by humans but also by machines and programs [6]. Meanings of terms (or concepts) in web resources are defined unambiguously using ontologies stated at the "annotation stage", which are written using SW languages, to represent conceptualization of application domains. Existing SW languages are all logic based (e.g., RDF and OWL are based on description logic [DL], and ontologies [11] are based on first order logic), and representation of meanings is expressed as logical sentences, and reasoning is done using logical inference. The SW languages are inadequate in specifying the semantics of probabilistic information. The final FSIRS web presentation can be classified into 6 different reporting categories, which will cover Pathogen Profiling, PMP Modeling Analysis, Microbial Risk Assessment, Outbreak Simulation Prediction, Rules, and Outbreak Trace Data Analysis as listed in **Fig. 3**. All the processed data in this system could be processed and presented for public reasoning with confidence, "trust" and "proof" with the help of dynamic advanced query algorithm "heuristic engine" (**Fig. 3**). As depicted in **Fig. 1** and **Fig. 3**, this final FSIRS web presentation would consist of six modules:

Profiling Module: an expert, easily viewed annotated system, which makes formulation recommendations based on experimental data that include biomarker detection for identification; standardized pulsed-field gel electrophoresis (PFGE)-based molecular subtyping; and molecular profiling based on new technologies, such as microarray, biometrics, and next-generation sequencing technologies.

PMP Modeling Module: a back propagation neural network, which predicts dissolution rate of the recommended formulation using the mapping between formulation parameters and dissolution rates learned from samples of lab test data. In order to increase accuracy in this module, each model will consist of two sets of supporting prediction modeling: one predicted through Bayesian Network Modeling and the other obtained from actual laboratory experiments.

Outbreak Trace Data Analysis Module: a database-based tracking system for surveillance.

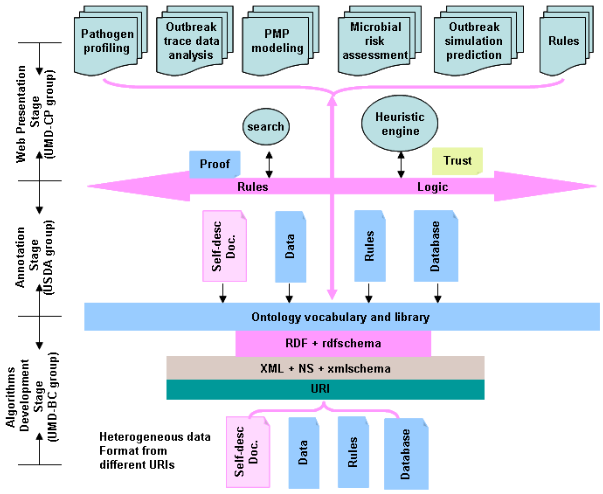
Pathogen profiling
Outbreak trace data analysis
PMP modeling
Microbial risk assessment
Outbreak simulation prediction
Rules
Web Presentation Stage (UMD-CP group)
search
Heuristic engine
Proof
Trust
Rules
Logic
Annotation Stage (USDA group)
Self-desc Doc.
Data
Rules
Database
Ontology vocabulary and library
RDF + rdfschema
XML + NS + xmlschema
URI
Algorithms Development Stage (UMD-BC group)
Heterogeneous data Format from different URIs
Self-desc Doc.
Data
Rules
Database

Figure 3: Knowledge based translation framework and the flow chart of research implementation plan (modified from Tim Berners Lee's vision of the Semantic).

Outbreak Simulation Prediction Module: based on simulation software to allow regulators and industry users to adjust variance factors when the predicted performance is not acceptable.

Risk Assessment Module: food-borne pathogen risk assessment is the quantitative or qualitative value of risk related to public health and the estimated potential loss through prediction modeling.

Government Response, Regulations Module: a regulation-based expert system.

CONCLUDING REMARKS

This paper has discussed a solution to a persistent problem facing current semantic web ontologies, namely, their inability of representing and reasoning upon uncertain relations and inputs. In particular, treating probabilistically enhanced OWL ontologies as probabilistic resources, the concept mapping between ontologies can be accomplished as evidential reasoning using the developed reasoning methods. Once this system is created, it will cover all important aspects of food-borne pathogens and cost less [20] considering that resources are routinely underfunded, technically “isolated”, or less comprehensive.

Robust, flexible, and extensible intelligent systems can be built in which both logical and probabilistic data of enormous quantity and variety in the web can be understood, utilized, and exchanged; knowledge can be continuously learned, integrated, and updated; and new and more complex problems can be solved. This means that food-borne pathogens are more likely to be detected and investigated comprehensively and sooner, whether they are in non-outbreak or outbreak status. Many new applications, which are not possible to accomplish at present, may emerge in the near future. These include, for example, data mining from data sources with different semantics, and probabilistic semantic integration and interoperation of software systems.

This work is only a step toward bringing the Information and Communication Technologies into the food safety research area. Its success will likely inspire more research in this direction. Richer representation and information may be developed, and new methods (e.g., reasoning with distributions or other probabilistic relations embedded in food-borne pathogen surveillance data) may appear.

Literature Cited

1. Adams BM, Saithanu K., and Hardin JM (2006) A neural network approach to control charts with applications to health surveillance. Invited talk at the 2006 Joint Statistical Meeting
2. Anonymous (1999). Health People 2000: Status Report - Food Safety Objectives. Food and Drug Administration, Food Safety and Inspection Service, and Centers for Disease Control and Prevention. September, 1999
3. Baranyi J, Tamplin M (2004) ComBase: A Common Database on Microbial Responses to Food Environments. J. Food Prot. 67:1834-1840
4. Baranyi J (2006) Using the ComBase database and associated software tools to predict microbial responses to food environments. Food manufacturing Efficiency. 1:9-13
5. Bean NH, Griffin PM (1990) Food borne disease outbreaks in the United States, 1973-1987: pathogens, vehicles, and trends. J. Food Prot. 53: 804-817
6. Berners-Lee T, Hendler J. and Lassila O (2001) The Semantic Web, Scientific American, May 2001
7. Ciccarese P, WuE, Wong G., Ocana M, Konshita J, Ruttenberg A, Clark T (2008) The SWAN biomedical discourse ontology. J. Biomedical Informatics 41:739-751
8. Cooper GF, Dash DH, Levander JD, Wong WK, Hogan WR, Wagner MM (2006) Bayesian Methods for Diagnosing Outbreaks In: Wagner MM, Moore AW, Aryel RM (ed) Handbook of Biosurveillance, Academic Press, Boston
9. Doan AH, Madhavan J, Domingos P, Halevy H (2002) Learning to map between ontologies on the Semantic Web, In WWW2002, May 7–11, 2002, Honolulu, Hawaii, USA.
10. Giugno R, Lukasiewicz T (2002) P-SHOQ(D): a probabilistic extension of SHOQ(D) for probabilistic ontologies in the Semantic Web, INFSYS Research Report 1843-02-06, Wien, Austria, April.
11. Gruber TR (1993) A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition, 5:199 – 220
12. Jiang X, Wallstrom GL (2006) A Bayesian network for outbreak detection and prediction, proceedings of the 21st national conference on Artificial intelligence, p.1155-1160, July 16-20, 2006. Boston, Massachusetts
13. Jiang X, Wallstrom GL, Cooper GF, Wagner MM (2009) Bayesian prediction of an epidemic curve, Journal of Biomedical Informatics, 42:90-99
14. Juneja VK, Huang CA (2009) Predictive microbiology information portal (PMIP) with particular reference to the USDA-pathogen modeling program (PMP), Blackwell Publisher (Accepted, Book Chapter).

15. Lacher MS, Groh G (2001) Facilitating the exchange of explicit knowledge through ontology mappings. In the proceeding of 14th Int. Florida A.I. Research Society Conf., pages 305–309. AAAI Press
16. Mansmann U (2005) Genomic profiling: Interplay between clinical epidemiology, bioinformatics and biostatistics. Methods Inf. Med. 44:454–460
17. McMeekin TA, Baranyi J, Bowman J, Dalgaard P, Kirk M, Ross T, Schmid S, Zwietering MH (2006) Information systems in food safety management. Int. J. Food Microbiol. 112:181-194
18. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV (1999) Food-related illness and death in the United States. Emerg. Infect. Dis. 5:607-625
19. Mitra P, Noy NF, Jaiswal AR (2004) OMEN: A Probabilistic Ontology Mapping Tool. In Workshop on Meaning Coordination and Negotiation at the Third International Conference on the Semantic Web (ISWC-2004). Hisroshima, Japan
20. Peck M, Baranyi J, Belsten J (2003) Microbial database could cut costs. Food Manufacturer. June/2003
21. Prasad S, Peng Y, Finin T (2002) A tool for mapping between two ontologies using explicit information, In AAMAS ’02 Workshop on Ontologies and Agent Systems. Italy, July 2002
22. Ross T, Baranyi J, McMeekin TA (1999) Predictive Microbiology and Food Safety. In:Robinson R, Batt C, Patel P (ed) Encyclopaedia of Food Microbiology, Academic Press, Boston
23. Sheth A (2004) From Semantic Search & Integration to Analytics, Dagstuhl Seminar on Semantic Interoperability and Integration, September 19-24, 2004. http://www.dagstuhl.de/04391/Materials/
24. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV (2001) PulseNet, the molecular subtyping network for foodborne bacterial disease surveillance, United States, Emerging Infectious Diseases 7:382–389
25. Swaminathan P, Gerner-Smidt LKNg, Lukinmaa S, Kam KM, Rolando S, Gutierrez EP, Binsztein N (2006) Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases, Foodborne Pathogens and Diseases 3:36–50
26. Taylor MR, Batz M (2008) Harnessing knowledge to ensure food safety: opportunities to improve the nation's food safety information infrastructure. Gainesville, FL: Food Safety Research Consortium; (Report available on the FSRC website at http://www.thefsrc.org/FSII/)
27. Tamplin M, Baranyi J, Paoli G (2003) Software programs to increase the utility of predictive microbiology information. In:McKellar RC, Lu X (ed) Modelling Microbial responses in Foods, CRC, Boca Raton, Fla
28. Villaneuva-Rosales N, Dumontier M (2008) yOWL: An ontology-driven knowledge base for yeast biologists. J. Biomedical Informatics 41:779-789