Application of Answer Set Programming for Public Health Data Integration and Analysis

Monica L. Nogueira and Noel P. Greis
Center for Logistics and Digital Strategy, Kenan-Flagler Business School,
The University of North Carolina at Chapel Hill,
Kenan Center CB#3440, Chapel Hill, NC 27713 U.S.A.
{monica_nogueira, noel_greis}@unc.edu

Abstract*—Public health surveillance systems routinely process massive volumes of data to identify health adverse events affecting the general population. Surveillance and response to foodborne disease suffers from a number of systemic and other delays that hinder early detection and confirmation of emerging contamination situations. In this paper we develop an answer set programming (ASP) application to assist public health officials in detecting an emerging foodborne disease outbreak by integrating and analyzing in near real-time temporally, spatially and symptomatically diverse data. These data can be extracted from a large number of distinct information systems such as surveillance and laboratory reporting systems from health care providers, real-time complaint hotlines from consumers, and inspection reporting systems from regulatory agencies. We encode geographic ontologies in ASP to infer spatial relationships that may not be evident using traditional statistical tools. These technologies and ontologies have been implemented in a new informatics tool, the North Carolina Foodborne Events Data Integration and Analysis Tool (NCFEDA). The application was built to demonstrate the potential of situational awareness—created through real-time data fusion, analytics, visualization, and real-time communication—to reduce latency of response to foodborne disease outbreaks by North Carolina public health personnel.

Keywords-data integration; answer set programming; public health; food safety; ontology

I. Introduction

Even though the U.S. food supply is one of the safest in the world, each year thousands of foodborne illness cases still occur causing irreversible human harm and extensive economic damage [1, 2]. The total cost of food contamination in the U.S. alone was recently estimated to be $152 billion, including health and human welfare costs as well as economic damage to companies and entire industries [3]. Surveillance and response to foodborne disease suffers from a number of systemic and other delays that hinder early detection and confirmation of emerging contamination situations. At the onset of an outbreak it is often impossible to link isolated events that may be related to other events reported by the same or by other data sources. In addition, it is often difficult to link events that may have a spatial relationship that is not immediately apparent from the data. For example, records from patients of different counties within the same state will only be reviewed by their respective local health departments. Once distinct events are suspected to be related, public health officials create a cluster and look for confirmatory evidence as part of a lengthy investigatory process. Latencies in the process could be reduced by the earlier availability and synthesis of other confirmatory information, including spatial information, often outside formal public health channels, including information from private companies and consumers [4].

In practice, local public health departments are usually the first to pick up the signals of foodborne disease. These signals may correspond to reports of illness generated by different types of events, e.g. (E1) a patient with symptoms of gastro-intestinal distress seeking medical attention at a hospital emergency room or a patient’s visit to a private physician’s office; (E2) laboratory test results for an ill patient which confirm a causative pathogen; and (E3) a cluster of ill patients due to a common pathogen. Routinely, a state’s syndromic surveillance system collects data from local health care providers about events of type (E1), (E2) and (E3) on a continuous basis, reporting them to the Centers for Disease Control and Prevention (CDC).

However, other events can signal an emerging foodborne disease outbreak. Many public health authorities and food industry operators, e.g. food manufacturers and grocery stores, maintain complaint hotlines (E4) where consumers report foodborne illness or a suspected adulterated food product. Consumer complaints made directly to public agency hotlines, e.g. local health departments (LHDs), state departments of agriculture or departments of environment and natural resources, are officially recorded and may lead to an investigation, at the discretion of the collecting agency.

A public food recall notification (E5) is another important signal which may be related to existing illness cases. Food manufacturers may voluntarily initiate the recall of one of their food products due to positive test results for foodborne pathogens, unintentional adulteration, mislabeling, and the presence of an allergen or hazardous material in the food product. Recalls may also be advised by authorities after routine inspections and testing conducted by the U.S. Department of Agriculture (USDA), U.S. Food and Drug Administration (FDA), and state agencies.

Food facility inspection reports (E6), which list violations to the food code applicable to such facilities, provide another signal that may inform and help identify the root cause of a contamination situation. Evaluation of the type, severity, and other characteristics of past code violations for a specific facility and the product(s) it manufactures could help link such operations as a probable source of contamination.

Microblogging and social media networks, i.e. Twitter or Facebook, are non-standard data sources that hold the potential, yet to be realized, to provide real-time information about emerging food contamination situations. Bloggers posting microblog messages on a social media network about illness after eating a certain food product or at a particular restaurant can provide timely indication about an emerging problem, referred to as type E7 event.

This paper expands on the work reported in [5], providing more detail and presenting additional work conducted since that paper was written. Our contributions with respect to rule-based event modeling are to: (1) extract relevant information from unstructured text, i.e. web-based recall notifications, to generate events that trigger our rule-based inference engine to "reason" about what it knows in light of the new information encoded by this event; (2) semantically link different types of events by employing (simple) ontologies for food, U.S. geographic regions, North Carolina counties, and foodborne diseases; (3) implement a rule-based inference engine using the Answer Set Programming (ASP) paradigm to identify emerging foodborne disease outbreaks; and (4) reduce latency in outbreak detection by identifying emerging outbreaks when the number of cases falls below the statistical threshold.

The paper is organized as follows. Section 2 discusses the motivation and challenges to represent the food safety domain using the ASP paradigm. Section 3 provides an overview of related work. Section 4 presents our rule-based event model and describes the ASP inference engine developed for our application. An illustrative example of the domain is shown in Section 5. Conclusions and future research directions are discussed in Section 6.

II. MOTIVATION AND CHALLENGES

Data associated with event types E1–E7 described above are collected by separate information systems and maintained and managed by distinct governmental agencies. Thus, in responding to the twin challenges of early detection of and rapid response to emerging outbreak situations, a central problem is how to access, process and interpret more events more quickly, thus reducing their time, scale, and scope. Framing the problem of outbreak detection as a complex event addresses a major failure of current surveillance methods. Current syndromic surveillance systems utilize statistics-based cumulative sum algorithms, i.e. CUSUM, to detect increases in illness reporting numbers and to determine that a foodborne disease outbreak may be emerging or is on-going. Alerts are normally generated by the system when the number of illness cases assigned to a certain syndrome, e.g. fever, respiratory, or gastro-intestinal distress, exceeds the threshold determined for that particular syndrome for the geographic area originating these events, i.e. county, and the local population baseline. These alerts are typically based solely on reported illness cases, type E1 to E3 events, above. Consideration of event types E4 to E7 aids in the detection of emerging outbreaks before they are sufficiently advanced to rise above the threshold of traditional CUSUM statistical methods.

A recent large foodborne outbreak in Europe provides a clear example of the challenges faced by public health officials worldwide. In May 2011 a new strain of the *Escherichia coli* bacteria, known as *E. coli O104:H4*, sickened more than 3,200 people and caused 36 deaths, mostly in Germany, but the source of contamination has not being confirmed at the time of this writing, on early June 2011 [6, 7]. Initial blame was attributed to raw tomatoes, lettuce and cucumbers but after tests returned with negative results, investigations moved to a new candidate – bean sprouts – with preliminary results still inconclusive. European farmers suffered economic losses amounting to hundreds of millions of euros. This led the European Union Farm Commissioner to propose a €150-million compensation package to aid farmers across Europe, which is being considered insufficient as it corresponds to just one-third of the estimated losses.

To better understand the investigatory processes for outbreak detection, consider the following situation. Two individuals in different but adjoining counties experience severe gastro-intestinal ulceration (GIU) symptoms after eating at their favorite local restaurant chain and seek medical attention at the emergency room of their local hospitals. Two separate illness reports are recorded and entered into the health care system to be reported to the state's public health syndromic surveillance system. First, if the number of reported GIU and foodborne-related cases does not exceed the corresponding threshold for GIU syndrome in the county, then no alert will be generated by the CUSUM algorithm and detection of an emerging situation will be delayed. Second, because CUSUM does not recognize adjoining counties even though the two individuals were made ill but the same pathogen/food no alert will be generated. However, consider that another person falls ill after eating at a different branch of the same food chain and calls a consumer complaint hotline to make a report. Currently, this event will be registered in the receiving agency's database but not automatically passed along to public health syndromic surveillance systems. Consider that another person, also ill after eating at that chain, reports the illness a message on a personal blog or preferred online discussion board. Both these events occur "under the radar" of public health and are not currently picked up as evidence of a possible emerging contamination situation.

III. RELATED WORK

The food safety domain is complex, multifaceted and dynamic. Recent high profile food contamination events have led to a surge of new regulations worldwide that provide food safety agencies with broader authority to enable (1) more stringent monitoring of the food supply chain; (2) data sharing among stakeholders; and (3) shifting from a remediation approach to preventive scientific-based risk analysis. But a widely accepted standard for tracking and tracing of food "from farm to fork" is still lacking, as well as computational models and analytical tools that will allow early recognition of emerging issues by fusing data from diverse sources and rapid response to adverse events. Work toward systems to improve identification of emerging food contamination problems is underway both in the U.S. and abroad. A review and classification of existing methods and reactive systems are provided in [8], including:

a) Early warning systems/networks monitor hazards in the food production chain through a centralized database which serves as a platform to communicate with and alert member agencies of emerging hazards, e.g. European Union's Rapid Alert System for Food and Feed (RASFF), European and U.S. Centers for Disease Control and Prevention (ECDC and CDC), International Food Safety Authorities Network (INFOSAN), etc.;

b) Combinatorial data systems monitor food safety data combined with other relevant data sources and advanced algorithms and technologies to identify and analyze associated hazards and risks in the food supply, e.g. Scientific Information and Expertise for Policy Support in Europe (SINAPSE), ECDC Communicable Disease Threat Reports (CDTRs), USDA Center for Emerging Issues (CEI) of the Animal and Plant Health Inspection Service (APHIS);

c) Retrospective analysis of reactive systems evaluate data collected by systems of type *a)* and *b)* above to detect trends that may be associated with the development of food safety issues, e.g. RASFF annual reports, European Food and Veterinary Office (FVO), food inspection agencies;

d) Proactive methods involve predictive risk assessment for hazard identification within the food production process and methods for routinely measuring production outcomes, e.g. Hazard Analysis Critical Control Points (HACCP); and

e) Vulnerability assessment focuses on the points within the food supply where hazards may be introduced and identification of indicators and countermeasures against potential hazards, e.g. U.S. CARVER-shock methodology.

The application developed and implemented as part of this research is both a combinatorial and retrospective analysis system which provides increased situational awareness to North Carolina's public health officials. Other related work that seeks to identify food safety issues spans different areas. An integrated information system for foodborne disease and food safety focused on pathogen data is discussed in [9]. The system proposed is ontology-driven, and utilizes the semantic web and heterogeneous data sources. Food allergens lead to a large number of food recalls each year as they pose a threat to people's health. Semantic web technologies, i.e. food allergen ontology, are discussed in [10] as bridges that will help promote data sharing and integration among food allergen databases. A new methodology described in [11] utilizes three data mining methods for recognizing patterns linking specific foodborne disease outbreaks with food products and consumption locations. An example of research that addresses the need for methods to trace food products "from farm to fork" can be found in [12]. A comprehensive review of existing methods to select indicators to help identify emerging hazards in the food supply chain appears in [13].

IV. Application Modeling

Although public health electronic surveillance systems collect and analyze daily massive volumes of data corresponding to patient visits to healthcare providers, it may take days or weeks to detect an emerging foodborne disease outbreak unless it is temporally and spatially localized, e.g. several people attending a banquet fall ill within a few hours and seek medical care at medical facilities located in neighboring geographical regions, e.g. counties. As discussed above, integration of data available to other food safety stakeholders, i.e. governmental agencies, private sector, and consumers, is critical to improve situational awareness and reduce such latencies. In this section we tackle the challenge of achieving data integration by modeling semantically heterogeneous data obtained from food-related events as the first step in developing our food safety application.

A. Event Model

An *event* is defined as the acquisition of a piece of information that is significant within a specific domain of interest to the application. In this application the domain of interest is food safety. We distinguish between two different types of events: simple events and materialized complex events. Simple events include both *atomic* events and *molecular* events. Atomic events have a distinct spatio-temporal identity, i.e. they take place at a particular place and time that is relevant to the determination of the complex event. An example of an atomic event would be a single reported case of gastrointestinal illness, an FDA recall, or a consumer complaint. Molecular events can be thought of as atomic events that are "linked together" by evidence, for example that are joined through previous evidence or by public health experts outside the system. Molecular events could include a confirmed cluster of two or more *Salmonella* cases as determined by DNA fingerprinting. An *event stream* is defined as the sequence of simple events received by the complex event processing (CEP) engine that are assigned a timestamp from a discrete ordered time domain and a geostamp consisting of a longitude and latitude geocode. An atomic event has a single timestamp and geostamp; molecular events may have multiple timestamps and geostamps.

A *complex* or materialized event is an event that is inferred by the engine's evaluation of the occurrence of other simple events. For example, in our application the materialized event is a foodborne disease outbreak.

B. Semantic Model

Our representation allows for incomplete information which is indicated by a unique reserved symbol of the representation language. Sparse data are an inherent characteristic of the problem, since one of our goals is to detect outbreaks when the number of illness cases has not yet exceeded thresholds employed by traditional statistical methods. The semantic model for the food safety domain is presented in Fig. 1. The events of interest are described by the following concepts.

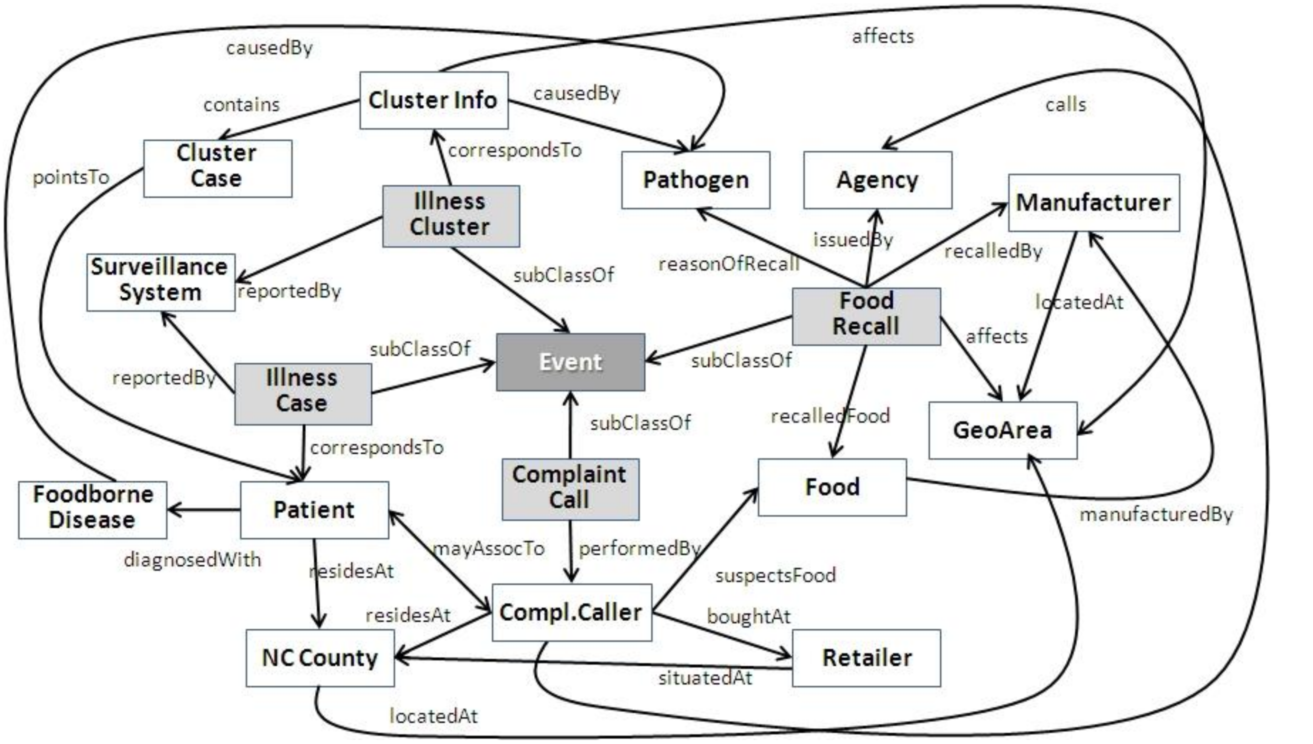


Figure 1. Semantic Model.

A patient illness case record corresponding to event type E1 contains information that uniquely identifies a patient in the record generating system, e.g. patient identification code and syndromic surveillance system; patient's county of residence; time and date of the visit to health care provider; syndrome or diagnosis assigned by attending physician, e.g. salmonellosis; and the disease-causing pathogen. This record will be updated to confirm the pathogen identified by a laboratory test when an event of type E2 corresponding to this patient enters the system. A simple ontology of foodborne diseases and related syndromes is employed to enable the semantic link of diagnosis data and pathogen data across different types of events. In this work, an ontology is a formal explicit description of concepts and individual instances of these concepts, which together constitute a knowledge base [14].

By definition, an illness cluster is formed by a number of patients with a common diagnosis caused by the same pathogen (as identified by laboratory test results or other causal links). The cluster patient with the earliest disease onset date is referred to as "patient#1." Events of type E3 are represented by two different types of records: (a) a cluster record provides information that uniquely identifies a specific cluster; and (b) a cluster illness case record contains information about a specific patient that is included in the cluster defined by a cluster record. A cluster record contains a unique identification code, the disease-causing pathogen, number of counties affected by the outbreak, number of patients in the cluster, unique identification code of patient#1, and date of patient#1 visit to a health care provider. A cluster illness case record contains the cluster unique identification code, the unique identification code of the patient it represents, and the patient's county of residence. A cluster illness case record acts as a pointer to the more complete patient illness case record of the corresponding patient.

A consumer complaint call, an event of type E4, is represented by three types of records. A complaint caller record provides a unique call identification code, date and time of call, and information about the caller, e.g. caller's state and county of residence; type of illness codified using the responding agency's medical code; and number of people that fell ill because of the product. A complaint food operator record contains information about the manufacturer and retailer the caller has complained about. A complaint food product record lists the food product as described by the caller and its corresponding FDA food code, date of manufacturing, and other information. A food ontology semantically links recalled food products to those implicated by consumer complaint calls. This ontology differs from existing food ontologies, e.g. [15], in that it closely follows the food categories as determined by FDA's product code which is utilized by industry to codify food shipments.

A recall notification, event of type E5, is represented by two types of records. A recall record contains a unique identification code for this event, the agency issuing the recall, date and time of its release, recalling company, recalled food product, reason for the recall, e.g. mislabeling, presence of allergen, or a specific pathogen, e.g. *Salmonella*, number of illnesses if known, and number of geographic areas affected. U.S. geographic areas are defined by a simple ontology which includes all U.S. states and regions as defined by the U.S. Census Bureau. An associated recall area record is created for each geographic area affected by the recall.

C. ASP Rule-Based Inference Engine

In this work, we use a form of declarative programming – Answer Set Programming (ASP) [16], to represent the rule-based CEP of the food safety domain and to search for/detect emerging outbreaks and other information of interest to public health officials. ASP has been applied to industrial problems but, to the best of our knowledge, it has not been used in food safety applications before.

The ASP paradigm is based on the stable models/answer sets semantics of logic programs [17, 18] and has been shown to be a powerful methodology for knowledge representation, including the representation of defaults and multiple interesting aspects of reasoning about actions and their effects, as well as being particularly useful to solve difficult search problems. In the ASP methodology, search problems are reduced to the computation of the stable models of the problem. Several ASP solvers – programs that generate the stable models of a given problem encoded in the ASP formalism – have been implemented, e.g. ASSAT [19], Clasp [20], Cmodels [21] DLV [22], GnT [23], nomore++ [24], Pbmodels [25], Smodels [26], etc. In what follows we provide the basic syntactic constructs and the intuitive semantics of the ASP language used in this work. A complete formal specification of the syntax and semantics of the language can be found at [18, 22].

A signature Σ of the language contains constants, predicates, and function symbols. Terms and atoms are formed as is customary in first-order logic. A literal is either an atom (also called a positive literal) or an atom preceded by $\neg$ (classical or strong negation), a negative literal. Literals l and $\neg l$ are called contrary. Ground literals and terms are those not containing variables. A consistent set of literals do not contain contrary literals. The set of all ground literals is denoted by $lit(\Sigma)$. A rule is a statement of the form:

$$h_1 \vee ... \vee h_k \leftarrow l_1, \dots, l_m, \mathit{not}\ l_{m+1}, \dots, \mathit{not}\ l_n. \tag{1}$$

where h_i's and l_i's are ground literals, *not* is a logical connective called default negation or negation as failure, and symbol $\vee$ corresponds to the disjunction operator. The head of the rule is the part of the statement to the left of symbol $\leftarrow$, while the body of the rule is the part on its right side. Intuitively, the rule meaning is that if a reasoner believes $\{l_1, \dots, l_m\}$ and has no reason to believe $\{l_{m+1}, \dots, l_n\}$, then it must believe one of the h_i's. If the head of the rule is substituted by the falsity symbol $\perp$ then the rule is called a constraint. The intuitive meaning of a constraint is that its body must not be satisfied. Rules with variables are used as a short hand for the sets of their ground instantiations. Variables are denoted by capital letters. An ASP program is a pair of $\{\Sigma, \Pi\}$, where Σ is a signature and Π is a set of rules over Σ, but usually the signature is defined implicitly and programs are only denoted by Π. A stable model (or answer set) of a program Π is one of the possible sets of literals of its logical consequences under the stable model/ answer set semantics.

D. ASP Program Encoding

Our encoding – the set of rules of program Π – contains roughly 100 rules, while event records (in ASP, rules with an empty body, also called "facts") for experiments are in the hundreds. We use the DLV system [22] as our ASP solver. To illustrate the ASP methodology we show below a few (simplified) rules used by our engine to detect emerging clusters. Rule (2) means that if neighboring counties A and B reported a small number of cases of food-related illnesses, due to pathogen P and/or syndrome S, this constitutes evidence for the engine to create a suspected cluster with case records – generated by rules of form (3) – from A and B with illness P and/or S. Then, an emerging outbreak affecting A and B, due to P, is computed by rule (4).

$$\begin{aligned}&\mathit{suspcluster}(A,B,P,S) \leftarrow\\ &\qquad \mathit{neighbors}(A,B),\\ &\qquad \mathit{min\_reached}(A,P,S),\\ &\qquad \mathit{min\_reached}(B,P,S).\end{aligned} \tag{2}$$

$$\begin{aligned}&\mathit{suspcluster\_illness}(A,B,Id,P,A) \leftarrow\\ &\qquad \mathit{suspcluster}(A,B,P,S),\\ &\qquad P\ != S,\\ &\qquad \mathit{patient\_illness}(Id,H,M,AmPm,Day,Mon,Y,A,Sys,P).\end{aligned} \tag{3}$$

$$\begin{aligned}&\mathit{susp\_outbreak}(A,B,P) \leftarrow\\ &\qquad \mathit{suspcluster}(A,B,P,\_).\end{aligned} \tag{4}$$

To assist public health officials with the task of identifying the food source of the contamination (e.g. fresh vegetable) and its manufacturer (e.g. MyVeggies), our engine seeks to connect a suspected cluster of illnesses to an existing public food recall notice. With rule (5), the engine will link a suspected cluster, occurring in neighboring counties A and B, to a food recall R, if the same pathogen P causing the cluster's illnesses is the reason for R; the area L affected by R encompasses these counties, e.g. L corresponds to the state where A and B are located, and the recall has been issued not too long ago, i.e. within a timeframe that makes the food recalled a probable candidate as the source of contamination for this cluster. By linking the suspected cluster to a food recall, rule (5) allows the inference of the food product causing the contamination, F, and its manufacturer, M.

$$\begin{aligned}&\mathit{susprecall}(R,A,B,F,M,L) \leftarrow\\ &\qquad \mathit{suspcluster}(A,B,P,S),\\ &\qquad \mathit{recall}(R,Mon,Y\ M,F,N,P,\ \_),\\ &\qquad \mathit{susprecall\_areaOK}(R,L),\\ &\qquad \mathit{susprecall\_dateOK}(R,A,B,Mon,Y).\end{aligned} \tag{5}$$

It may be possible for a given cluster to be linked to more than one food recall, since it not uncommon to have more than one food product contaminated with a given pathogen at the same time. This is particularly true in ingredient-based outbreaks, where an ingredient, e.g. peanut butter paste, is used in multiple products, e.g. soups and ice cream, which leads to multiple recall notices by various manufacturers. Facts (6)–(11) describe the recall of two different food products, e.g. R1: meat and R2: chicken, due to the same pathogen, *Salmonella*, occurring at the same time, March 2011, with no (zero) officially confirmed illnesses. Recall R1 affects only 3 states, while R2 covers a much larger area, i.e. the whole country.

```
recall(r1,mar,2011,meatfarm,meat,3,salmonella,0).   (6)
recall_area(r1,new_york).                           (7)
recall_area(r1,new_jersey).                         (8)
recall_area(r1,north_carolina).                     (9)

recall(r2,mar,2011,mama,chicken,1,salmonella,0).   (10)
recall_area(r2,nationwide).                        (11)
```

Information about the food product distribution area is frequently provided in food recall notices but is not always specific. A recall may indicate one or more states where the food product has been distributed; cite a whole region, e.g. U.S. Midwest; a county or city, e.g. New York City; or indicate it was distributed “nationwide”. In situations where multiple recalls are linked to a cluster due to a causing pathogen, and time and geographical coverage, it may be more efficient for public health officials to focus on recalls whose product distribution is closely related to the cluster’s geographical area. Thus, if both recalls R1 and R2 are connected to a local cluster in New York, New Jersey, or North Carolina, these states’ public health officials may give higher priority to investigating “MeatFarm’s meat” products as the source of contamination instead of “Mama’s chicken” products.

We allow users to select if they want to be informed by the engine of all recalls connected to a given suspected cluster, or to be informed only of those recalls that are inferred as “preferred recalls” by the engine due to a more specific geographic link to the cluster. It is ease to express such parametric computations using ASP. In our program, this is accomplished by introducing an atom – prefrecallON – to the set of (truthful) facts to be evaluated by the engine. Preferred recalls can then be inferred by rule (12). Intuitively, this rule says that given two distinct recalls, R1 and R2, already connected to a cluster formed by counties A and B, where area L1 affected by R1 is a subregion of area L2 affected by R2, the engine concludes that R1 is a “preferred recall,” of interest to public health officials, unless it can prove – using rule (13) – that there exists another recall, a certain distinct R3, covering a geographical area L3 which is a subregion of L1, and thus more specific to cluster A, B. Notice that if users are not interested that the engine generates “preferred recall” information, they may select not to add atom prefrecallON to the program. The absence of this fact falsifies rules (12) and (13), and no preferred recalls are generated by the engine.

```
pref_susprecall(R1,A,B,F1,M1, L1) ←                (12)
    prefrecallON,
    susprecall(R1,A,B,F1,M1,L1),
    susprecall(R2,A,B, F2,M2, L2),
    subregion(L1,L2),
    R1 != R2,
    not other_more_specif(A,B,L1, R2).

other_more_specif(A,B,L1,R2) ←                     (13)
    prefrecallON,
    susprecall(R2,A,B,_,_,L2),
    susprecall(R3,A,B,_,_,L3),
    subregion(L3, L1),
    R2 != R3.
```

The highly expressive power of ASP facilitates the creation of knowledge hierarchies, organization and inference of concepts which are primary requirements for the development of ontologies. We take advantage of the language’s expressiveness to represent three (currently simple) ontologies in the following areas: (a) foodborne disease; (b) food; (c) U.S. geographical divisions and North Carolina political divisions. Due to space limitations, only the geographical ontologies are briefly described next.

E. *Geographical Ontologies in ASP*

The concepts represented in our geographical ontologies are the standards: country, regions, regional divisions, states, territories, counties, and cities. Fig. 2 illustrates the main concepts and relations of these ontologies. Given that public health officials in the U.S. are mostly interested in solving foodborne disease outbreaks occurring in the country, our data are limited to the United States. In particular, the application developed is to assist North Carolina authorities

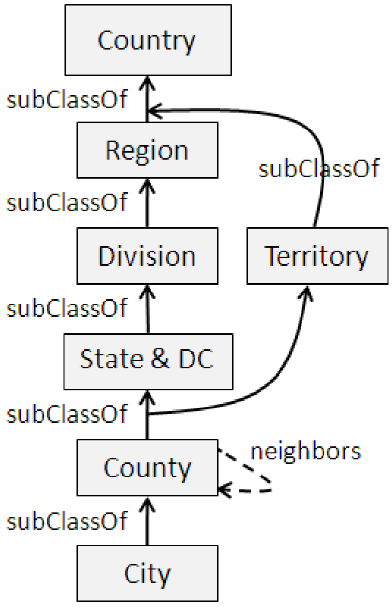


Figure 2. Geographical Ontology.

and thus, at present, county and city data are limited to this state. For implementation purposes, we distinguish the U.S. Ontology from the North Carolina (NC) Ontology, given that the latter includes a neighboring relation among NC counties. The main goals of the geographical ontologies are to enable the engine to infer: (1) preferred food recalls based on their geographic specificity to the state of North Carolina; and (2) clusters of foodborne disease covering a number of neighboring NC counties.

States are represented by facts of type (14), which include the state name, abbreviation, and regional division. We encode 4 U.S. regions which are divided into 9 regional divisions, as adopted by the U.S. Census Bureau. The state of North Carolina is located in the South Atlantic division of the South region, as encoded by (14)–(16). Territories are similarly represented. Part of the food products' distribution information is represented by facts (17) – (21), where "nationwide" is an example of a standard way food recall notices indicate a product has been distributed to the whole country. Rules (22)–(25) are used by rules (12) and (13) to prove that a given recall area, namely L1, is a subregion of another larger area, L2, and therefore information about the recall on-going in L1 is preferred than that on-going in L2.

$$us\_state(north\_carolina, nc, south\_atlantic). \tag{14}$$
$$division(south\_atlantic, south). \tag{15}$$

$$region(R) \leftarrow division(\_, R). \tag{16}$$

$$is\_us\_distr(nationwide). \tag{17}$$
$$is\_us\_distr(R) \leftarrow region(R). \tag{18}$$
$$is\_us\_distr(R) \leftarrow division(R,\_). \tag{19}$$
$$is\_us\_distr(S) \leftarrow us\_state(S, \_, \_). \tag{20}$$
$$is\_us\_distr(S) \leftarrow us\_state(\_, S, \_). \tag{21}$$

$$subregion\ (R, nationwide) \leftarrow region(R). \tag{22}$$
$$subregion(S,R) \leftarrow division(S,R). \tag{23}$$
$$subregion(S,R) \leftarrow us\_state(S, \_, R). \tag{24}$$
$$subregion\ (S,R) \leftarrow us\_state(\_, S, R). \tag{25}$$

To infer new suspected clusters of foodborne illness, rule (2) tries to prove that there are two neighboring counties in the state with at least a minimum number of reported illness cases due to the same pathogen or with a generic GIU diagnosis that may be caused by food contamination. As shown in rule (3), case records inform the patients' county of residence. The simple NC ontology encodes 100 NC counties, e.g. fact (26), more than 100 neighbor relations for these counties, e.g. facts (27)–(30), and around 850 cities distributed throughout these counties, e.g. (31).

$$nc\_county(orange). \tag{26}$$

$$nc\_neighbors(orange, alamance). \tag{27}$$
$$nc\_neighbors(orange, caswell). \tag{28}$$
$$nc\_neighbors(orange, chatham). \tag{29}$$
$$nc\_neighbors(orange, person). \tag{30}$$

$$nc\_city(chapel\_hill, orange). \tag{31}$$

F. *Evidence Set and Event Evidence Indicator*

The set of linked events that provide evidence of the materializing of a complex event is called the *Evidence Set*. An evidence set is associated with a degree of uncertainty as to whether an emerging outbreak event will materialize based on the information in the event data. Rule (32) below exemplifies the computation of elements of the evidence set. Intuitively, the rule meaning is that the engine will conclude that there is evidence that a complaint call C, from county T implicating a food product F1 of type FC per the FDA Code, is connected to a materialized cluster, of illness P affecting neighboring counties A and B, if this call can be linked to an existing recall R of food F2 manufactured by M at location L, if food 2 is also of type FC.

$$\begin{aligned} &evidence(A,B,P,S,R,F2,M,L,F1,FC,T) \leftarrow \\ &\quad suspcluster(A,B,P,S), nccounty(T), \\ &\quad suspcomplaint(C,A,B,F1,FC,T), \\ &\quad susprecall(R,A,B,F2,M,L), type\_of(F2,FC). \end{aligned} \tag{32}$$

NCFEDA's engine computes a measure of the strength of the evidence supporting the conclusion of an emerging complex event through a ranking that ranges from 0, or no evidence, to a maximum of 7, highest evidence rating. The computation of the *Event Evidence Indicator* (EVI) is based on the number and strength of the relationships that connect the events in the evidence set and corresponds to the weighted summation of EVI components calculated for the subsets formed when linking pairs of different types of events. For example, we compute the EVI component for the set of all events corresponding to a suspected cluster and incoming recall notification.

V. Illustrative Application

The ASP rule-based inference engine was implemented in the North Carolina Foodborne Events Data Integration and Analysis Tool (NCFEDA) shown in Fig. 3. In this system, food-related events are received by the *Events Manager* which consists of two components: (1) a set of databases; and (2) the *Event Trigger Module*. These databases store all food-related events and geocoded datasets across all public and private sector stakeholders contributing to NCFEDA. The Event Manager monitors the databases for new incoming events that are evaluated by the Event Trigger Module for selecting possible triggers. As noted earlier, and in Fig. 3, triggers are events that could include a case related to a foodborne illness or a consumer complaint.

Web scraping techniques are utilized by NCFEDA to obtain information about recalls of food products issued by the FDA and USDA directly from their websites. Given that recall notifications are intended for a human audience, they are written in natural language which poses an additional challenge for the extraction of information. A *Semantics Module* enables the extraction of events from unstructured data by automatically parsing recall text and extraction of relevant information. We utilize a hybrid two-step method to perform the data extraction. The first step takes advantage of the metadata information contained in the FDA and

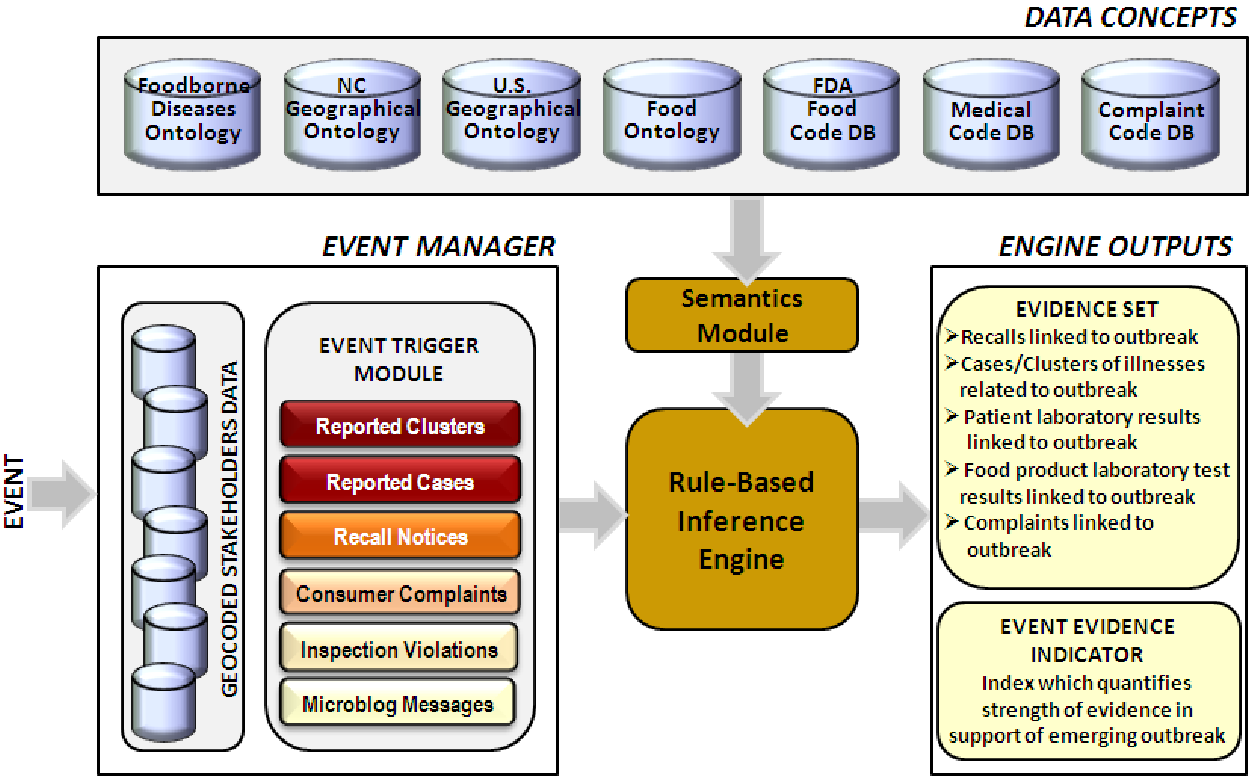

DATA CONCEPTS
Foodborne Diseases Ontology
NC Geographical Ontology
U.S. Geographical Ontology
Food Ontology
FDA Food Code DB
Medical Code DB
Complaint Code DB
EVENT MANAGER
EVENT
GEOCODED STAKEHOLDERS DATA
EVENT TRIGGER MODULE
Reported Clusters
Reported Cases
Recall Notices
Consumer Complaints
Inspection Violations
Microblog Messages
Semantics Module
Rule-Based Inference Engine
ENGINE OUTPUTS
EVIDENCE SET
➢Recalls linked to outbreak
➢Cases/Clusters of illnesses related to outbreak
➢Patient laboratory results linked to outbreak
➢Food product laboratory test results linked to outbreak
➢Complaints linked to outbreak
EVENT EVIDENCE INDICATOR
Index which quantifies strength of evidence in support of emerging outbreak

Figure 3. NCFEDA System Architecture.

USDA food recall webpages. Basic pattern matching of single terms is utilized to filter metadata fields of interest, e.g. date a recall was issued. A second step is required for information contained in the recall notification but not available as metadata, e.g. geographic areas where the food product has been distributed or number of confirmed illness cases attributed to food being recalled. Sequences of alternative patterns based on the ontologies developed are used to perform pattern matching and information extraction.

The Events Manager and Semantics Module work in conjunction with the *Rule-Based Inference Engine* to analyze incoming events and determine their relevance to the current situation(s) being monitored and/or other events previously stored in the system. NCFEDA databases must be maintained to remain up-to-date with available new information. The bulk of these data constitutes the history of food safety in North Carolina and will be used by the inference engine to support or refute possible conclusions regarding emerging food situations. Every new event arrival is analyzed and may or may not activate one or more of NCFEDA's inference engine rules. Using knowledge encoded as inference rules, trigger events are processed, together with other possible relevant events, by the Rule-based Inference Engine. The engine uses rules to "reason" about an existing situation as described by the known facts and encoded rules, and deduces relationships among events to determine whether there is an emerging outbreak event.

The screen shot shown in Fig. 4 illustrates the data integration and analysis capabilities of the application. The data sources are: 1) all illness cases records reported to the North Carolina Department of Public Health containing among other fields: the office visit date, probable diagnosis, and a patient's county of residence; 2) all recall notices of food products issued by the FDA stored as records containing: the recall issuing date, the product recalled, the company recalling the product, the cause for the recall – i.e. pathogen causing the contamination when available, and areas (states) where the product has been distributed; and 3) all consumer complaint calls implicating a food product including all of the following fields if available: date of the call, complainant county of residence, product implicated, retailer/manufacturer/food service provider implicated, complainant medical status (i.e. illness, hospitalization, etc.), diagnosis, and description of the complaint.

In the scenario illustrated, a new small cluster of salmonellosis cases has been detected. The application searches among incoming consumer complaint calls for any illnesses caused by *Salmonella* or any implicating food products susceptible to this pathogen. The search is narrowed down to a call reporting a hospitalization due to possible consumption of contaminated fruit. The application then searches among both incoming and previously active recall notices to determine whether any potentially relevant *Salmonella* recalls have been reported by FDA and, among those, whether any contaminated product has been shipped to North Carolina. The engine flags an incoming recall of a product to the state of North Carolina and previously restricted to three other states on the West Coast of the United States. The ontologies allow the recognition that cantaloupe is a fruit and that the pathogen causing this recall is also the same pathogen causing a reported illness and hospitalization, as reported by a consumer complaint call. Fig. 4 displays the location of all relevant events linked to this threat. When the user hovers the computer mouse over the map icons, detailed information about each reported case/complaint is displayed. The Event Evidence Indicator for the event is computed (EVI=5) and reported on the corner of the pop-up window along with a message relating the illness reports, recalls and consumer complaints, as well as the suspected food product (cantaloupe) and the pathogen (*Salmonella*).

VI. Conclusions and Future Directions

One of the central challenges in outbreak detection has been the inability to detect signals of an emerging outbreak from cases of illness that may not pass a statistical threshold because they are spatially and temporally dispersed—and for which there are considerable time lags. A primary contribution of this paper has been to frame the outbreak detection problem as a complex event where events include not only structured event data (e.g. case information) but also unstructured event data (e.g. recall or complaint data). We develop semantic models that are able to extract meaningful information from unstructured text data that can serve as event triggers. Using ontologies and rules we are able to discover semantic links between events that provide evidence of an emerging outbreak event. Identification of events that comprise the new suspected clusters of foodborne illness is accomplished using ASP. We successfully implemented these concepts in the NCFEDA prototype. This work is on-going and we are continuing to further develop the rule-based inference engine and we plan to explore the use of the OntoDLV language for our ontology representations [27].

Acknowledgment

This project is funded through a subcontract with the Institute for Homeland Security Solutions (IHSS). IHSS is a research consortium established to conduct applied research in the social and behavioral sciences. This material is based upon work supported by the U.S. Department of Homeland Security (DHS) under Contract No. HSHQDC-08-C-00100. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the official policy or position of DHS or of the U.S. Government. The study is approved for public release.

The authors also thank Brett Weed, Compliance and Preparedness Administrator of the Food and Drug Protection Division of the North Carolina Department of Agriculture and Consumer Services, and the members of the North Carolina Food Safety and Defense Task Force for their assistance with this research.

References

[1] E. Scallan, et al., “Foodborne illness acquired in the United States–Major pathogens,” Emerg Infect Dis, vol. 17(1), 2011, pp. 7–15.

[2] J.G. Morris, “How safe is our food?” Emerg Infect Dis, vol. 17(1), 2011, pp. 126–128.

[3] R.L. Scharff, “Health-Related Costs from Foodborne Illness in the United States.” The Produce Safety Project at Georgetown University, Washington, D.C., March 2010.

[4] N.P. Greis and M.L. Nogueira, Food Safety Emerging Public-Private Approaches: A perspective for local, state, and federal government leaders. IBM Center for The Business of Government, 2010.

[5] M.L. Nogueira and N.P. Greis, “Rule-Based Complex Event Processing for Food Safety and Public Health,” RuleML 2011 - Europe, N. Bassiliades et al., Eds., LNCS, vol. 6826. Springer, Heidelberg, 2011, pp. 376–383, in press.

[6] Associated Free Press, “Killer bacteria toll rises to 36.” AFP, 13 June 2011. Web. 13 June 2011.

[7] A. Cowell, “Germany Faces Criticism Over E. Coli Outbreak.” *NYTimes.com*. 7 June 2011. Web. 7 June 2011. <www.nytimes.com/2011/06/08/world/europe/08ecoli.html>.

[8] H.J.P. Marvin, et al., “A working procedure for identifying emerging food safety issues at an early stage: Implications for European and international risk management practices,” Food Control, vol. 20. Elsevier, 2009, pp. 345–356.

[9] X. Yan, et al., “From Ontology Selection and Semantic Web to an Integrated Information System for Food-borne Diseases and Food Safety,” Software Tools and Algorithms for Biological Systems, H.R. Arabnia and Q.-N. Tran, Eds., Advances in Experimental Medicine and Biology, vol. 696, 2011, pp. 741–750.

[10] S.M. Gendel, “Allergen databases and allergen semantics,” Regulatory Toxicology and Pharmacology, vol. 54, 2009, pp. S7–S10.

[11] M. Thakur, S. Olafsson, J.-S. Lee and C.R. Hurburgh, “Data Mining for recognizing patterns in foodborne disease outbreaks,” J Food Engineering, vol. 97, 2010, pp. 213–227.

[12] A. Regattieri, M. Gamberi and R. Manzini, “Tracebility of food products: general framework and experimental evidence,” J. Food Engineering, vol. 81, 2007, pp. 347–356.

[13] G.A. Kleter and H.J.P. Marvin, “Indicators of emerging hazards and risks to food safety,” Food and Chemical Toxology, vol. 47, 2009, pp. 1022–1039.

[14] N. Noy and D. McGuiness, “Ontology Development 101: A Guide to Creating Your First Ontology.” Technical Report SMI-2001-0880, Stanford University, 2001.

[15] J. Cantais, D. Dominguez, V. Gigante, L. Laera and V. Tamma, “An example of food ontology for diabetes control,” Proc. of the International Semantic Web Conference 2005 workshop on Ontology Patterns for the Semantic Web, Galway, Ireland, November 2005.

[16] V.W. Marek and M. Truszczynski, “The Logic Programming Paradigm: a 25-Year Perspective,” chap. Stable models and an alternative logic programming paradigm. Berlin: Springer Verlag, 1999, pp. 375–398.

[17] M. Gelfond and V. Lifschitz, “The stable model semantics for logic programming,” International Logic Programming Conference and Symposium, R. Kowalski and K. Bowen, Eds. MIT Press, 1988, pp. 1070–1080.

[18] M. Gelfond and V. Lifschitz, “Classical negation in logic programs and disjunctive databases,” New Generation Computing, vol. 9, 1991, pp. 365–385.

[19] F. Lin and Y. Zhao, “ASSAT: Computing answer sets of a logic program by SAT solvers,” Artificial Intelligence, vol. 157(1–2), 2004, pp. 115–137.

[20] M. Gebser, B. Kaufmann, A. Neumann and T. Schaub, "*clasp*: A Conflict-Driven Answer Set Solver," Proc. 9th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR'07), Baral, C., Brewka, G. and J. Schlipf, Eds., LNCS, vol. 4483, 2007, pp. 260–265.

[21] Y. Lierler, "Cmodels—SAT-based disjunctive answer set solver," Proc. 8th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR'05), Baral, C., Greco, G., Leone, N., and G. Terracina, Eds., LNAI vol. 3662, Italy, 2005, pp. 447–451.

[22] N. Leone, et al., "The DLV System," Proc. 8th European Conference on Logics in Artificial Intelligence (JELIA 02), Flesca, S., Greco, S., Ianni, G. and N. Leone, Eds., LNCS vol. 2424, Italy, September 2002, pp. 537–540.

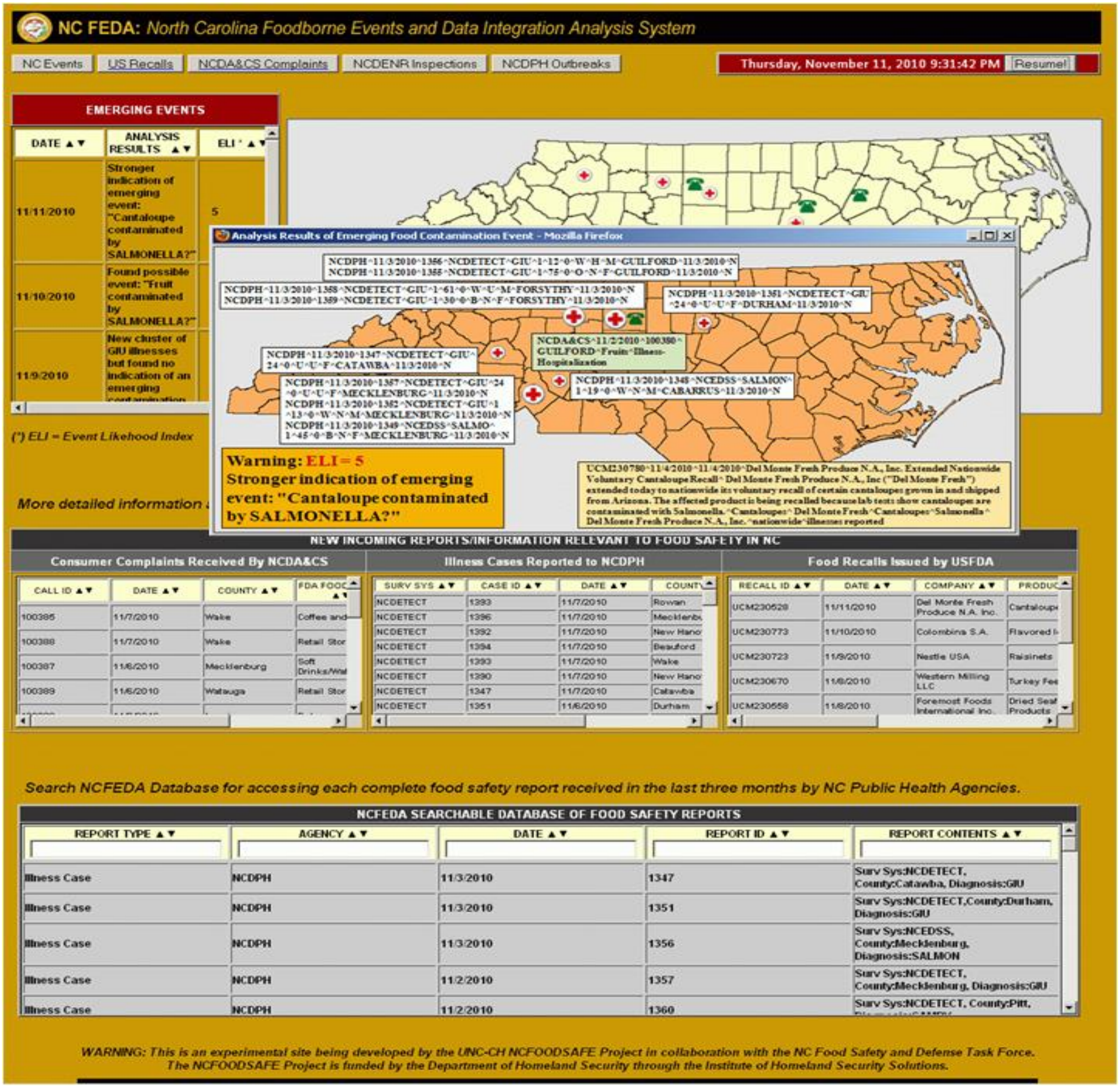
[23] T. Janhunen, I. Niemelä, D. Seipel, P. Simons and J.-H. You, "Unfolding Partiality and Disjunctions in Stable Model Semantics," ACM Transactions on Computational Logic, vol. 7(1), January 2006, pp. 1–37.

[24] C. Anger, M. Gebser, T. Linke, A. Neumann and T. Schaub, "The nomore++ System," Proc. 8th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR'05), Baral, C., Greco, G., Leone, N., and G. Terracina, Eds., LNAI vol. 3662, Italy, 2005, pp. 422–426.

[25] M. Truszczynski, "Predicate-calculus-based logics for modeling and solving search problems," ACM Transactions on Computational Logic, vol. 7 (1), 2006, pp. 38–83.

[26] I. Niemelä and P. Simons, "Logic-Based Artificial Intelligence," chap. Extending the Smodels System with Cardinality and Weight Constraints. Kluwer Academic Publishers, 2000, pp. 491–521.

[27] F. Ricca, et al., "OntoDLV: An ASP-based System for Enterprise Ontologies," Journal of Logic and Computation, vol. 19(4). Oxford University Press, 2008, pp. 643–670.


NC FEDA: North Carolina Foodborne Events and Data Integration Analysis System
NC Events
US Recalls
NCDA&CS Complaints
NCDENR Inspections
NCDPH Outbreaks
Thursday, November 11, 2010 9:31:42 PM
Resume!
Analysis Results of Emerging Food Contamination Event - Mozilla Firefox

EMERGING EVENTS

DATE ▲▼	ANALYSIS RESULTS ▲▼	ELI * ▲▼
11/11/2010	Stronger indication of emerging event: "Cantaloupe contaminated by SALMONELLA?"	5
11/10/2010	Found possible event: "Fruit contaminated by SALMONELLA?"	
11/9/2010	New cluster of GIU illnesses but found no indication of an emerging contamination	

(*) ELI = Event Likehood Index

More detailed information

NCDPH^11/3/2010^1356^NCDETECT^GIU^1^12^0^W^H^M^GUILFORD^11/3/2010^N
NCDPH^11/3/2010^1355^NCDETECT^GIU^1^75^0^O^N^F^GUILFORD^11/3/2010^N

NCDPH^11/3/2010^1358^NCDETECT^GIU^1^61^0^W^U^M^FORSYTHY^11/3/2010^N
NCDPH^11/3/2010^1359^NCDETECT^GIU^1^30^0^B^N^F^FORSYTHY^11/3/2010^N

NCDPH^11/3/2010^1351^NCDETECT^GIU^24^0^U^U^F^DURHAM^11/3/2010^N

NCDA&CS^11/2/2010^100380^GUILFORD^Fruit^Illness-Hospitalization

NCDPH^11/3/2010^1347^NCDETECT^GIU^24^0^U^U^F^CATAWBA^11/3/2010^N

NCDPH^11/3/2010^1357^NCDETECT^GIU^24^0^U^U^F^MECKLENBURG^11/3/2010^N
NCDPH^11/3/2010^1352^NCDETECT^GIU^1^13^0^W^N^M^MECKLENBURG^11/3/2010^N
NCDPH^11/3/2010^1349^NCEDSS^SALMO^1^45^0^B^N^F^MECKLENBURG^11/3/2010^N

NCDPH^11/3/2010^1348^NCEDSS^SALMON^1^19^0^W^N^M^CABARRUS^11/3/2010^N

Warning: ELI = 5
Stronger indication of emerging event: "Cantaloupe contaminated by SALMONELLA?"

UCM230780^11/4/2010^11/4/2010^Del Monte Fresh Produce N.A., Inc. Extended Nationwide Voluntary Cantaloupe Recall^ Del Monte Fresh Produce N.A., Inc ("Del Monte Fresh") extended today to nationwide its voluntary recall of certain cantaloupes grown in and shipped from Arizona. The affected product is being recalled because lab tests show cantaloupes are contaminated with Salmonella.^Cantaloupes^ Del Monte Fresh^Cantaloupes^Salmonella^ Del Monte Fresh Produce N.A., Inc.^nationwide^illnesses reported

NEW INCOMING REPORTS/INFORMATION RELEVANT TO FOOD SAFETY IN NC

Consumer Complaints Received By NCDA&CS

CALL ID ▲▼	DATE ▲▼	COUNTY ▲▼	FDA FOOD ▲▼
100385	11/7/2010	Wake	Coffee and
100388	11/7/2010	Wake	Retail Stor
100387	11/6/2010	Mecklenburg	Soft Drinks/Wat
100389	11/6/2010	Watauga	Retail Stor

Illness Cases Reported to NCDPH

SURV SYS ▲▼	CASE ID ▲▼	DATE ▲▼	COUNTY
NCDETECT	1393	11/7/2010	Rowan
NCDETECT	1396	11/7/2010	Mecklenbu
NCDETECT	1392	11/7/2010	New Hano
NCDETECT	1394	11/7/2010	Beauford
NCDETECT	1393	11/7/2010	Wake
NCDETECT	1390	11/7/2010	New Hano
NCDETECT	1347	11/7/2010	Catawba
NCDETECT	1351	11/6/2010	Durham

Food Recalls Issued by USFDA

RECALL ID ▲▼	DATE ▲▼	COMPANY ▲▼	PRODUC
UCM230528	11/11/2010	Del Monte Fresh Produce N.A. Inc.	Cantaloup
UCM230773	11/10/2010	Colombina S.A.	Flavored I
UCM230723	11/9/2010	Nestle USA	Raisinets
UCM230670	11/8/2010	Western Milling LLC	Turkey Fee
UCM230558	11/8/2010	Foremost Foods International Inc.	Dried Sea Products

Search NCFEDA Database for accessing each complete food safety report received in the last three months by NC Public Health Agencies.

NCFEDA SEARCHABLE DATABASE OF FOOD SAFETY REPORTS

REPORT TYPE ▲▼	AGENCY ▲▼	DATE ▲▼	REPORT ID ▲▼	REPORT CONTENTS ▲▼
Illness Case	NCDPH	11/3/2010	1347	Surv Sys:NCDETECT, County:Catawba, Diagnosis:GIU
Illness Case	NCDPH	11/3/2010	1351	Surv Sys:NCDETECT,County:Durham, Diagnosis:GIU
Illness Case	NCDPH	11/3/2010	1356	Surv Sys:NCEDSS, County:Mecklenburg, Diagnosis:SALMON
Illness Case	NCDPH	11/2/2010	1357	Surv Sys:NCDETECT, County:Mecklenburg, Diagnosis:GIU
Illness Case	NCDPH	11/2/2010	1360	Surv Sys:NCDETECT, County:Pitt,

WARNING: This is an experimental site being developed by the UNC-CH NCFOODSAFE Project in collaboration with the NC Food Safety and Defense Task Force. The NCFOODSAFE Project is funded by the Department of Homeland Security through the Institute of Homeland Security Solutions.

Figure 4. Screenshot of NC Events Page of NCFEDA Use Case Showing A Warning with EVI=5.