# WHITE PAPER

# IT Subcommittee Report: Design and Development of an Inspection Results Collection/Reporting System

## Introduction

This report captures the research and recommendations of the IT Subcommittee. The IT Subcommittee is charged with assessing the existing landscape, proposing a system and methodology, and assessing project costs.

### Subcommittee members

Subcommittee membership is open to all interested committee members. Initial members are the following:

- Darryl Booth
- Chirag Bhatt
- Janice Buchanon
- Bryan Chapman
- Phillip Leslie
- Angela Nardone
- Ernesto Nardone
- Todd Taylor

## Assessment of Existing Landscape

In this section we review several organizations and projects currently working to provide some form of inspection database standardization or integration of retail food safety inspection results from local health departments. These projects vary in their purpose and details, as we describe below.

### University of Maryland Project: “Digital Disclosure of a Nationally Standardized Database of Restaurant Food Safety Inspections”

Funded by the Sloan Foundation, this project created a national database of restaurant food safety inspection results. The researchers retrieve inspection results from state and local health department web sites and integrate the data into a single integrated database.

The inspection results are not standardized — all the differences in the violation definitions across jurisdictions are preserved in the underlying data without imposing any assumptions to apply a standard.

The database is searchable. For example, a user can search for all food temperature violations at any restaurant during a specific period of time.

Not all jurisdictions put their data online. Those that do so, provide the results in many different formats. Simply scraping and integrating the data into an integrated database requires customization for each jurisdiction.

### Yelp.com

Yelp is a for-profit entity that collects and curates user-generated reviews. In January 2013 Yelp announced a plan to incorporate health department food safety inspections into their results.

Central to their plan is a standardized inspection score format across jurisdictions. Yelp partnered with the Cities of San Francisco and New York to generate a common data schema that health departments may utilize in providing their inspection data to Yelp. The schema is called the Local Inspector Value-Entry Specification (LIVES).

The LIVES format does not enforce a standard definition of violations.

### HDScores.com

HDScores.com is a Maryland-based for-profit startup that aims to “scrape” health department web sites for inspection data. The main goal appears to be utilizing the data to generate sales leads for restaurant hygiene solution providers.

### CalCode Data Dictionary

The CalCode Data Dictionary, established in 2006, was created to provide a standardized data schema for California health departments to provide electronic records of retail food safety inspection results.

### Local Data Management Systems

There exist several companies that provide software and IT services to health departments, including an accounting of restaurant inspection results.

In addition, any number of health departments may have locally-developed systems.

### State-Wide Systems

Approximately sixteen states (plus Washington DC) have integrated inspection in a single database at the state level.[1]

In these instances, the inspections themselves may or may not be standardized across reporting health departments within the state.

## Proposed System and Recommended Methodology

[1] Alabama, Delaware, DC, Florida, Georgia, Iowa, Kansas, Louisiana, Mississippi, North Carolina, Oklahoma, Pennsylvania, Rhode Island, South Carolina, Tennessee, Vermont, and Virginia.

## Pull Over Push

A traditional approach to the challenge of standardized and consolidated retail food inspection results is to 1) publish a data dictionary; 2) advocate its adoption through CFP, NEHA, and others; and 3) rely upon state and local health departments to voluntarily standardize and “push” their inspection data according to the documented standard.

This proposition is not particularly attractive to local health departments. The motives to establish and maintain the data-flows from the local to the consolidated system just are not compelling. Without a national requirement and financial remuneration, this traditional approach is likely to start slowly, struggle, or even fail at the local level.

Why is it that one can search Google for “Restaurant Name, Restaurant Number” and find recent inspection results from the Boulder County Environmental Health web site? It’s through Google’s engineering and computing power. On a regular basis, Google scours the Internet, following web pages link-by-link, until it eventually discovers Boulder County’s inspection results and the “Restaurant” inspection history. Boulder County didn’t have to do anything… except to publish its inspection results to the web (a best practice) for Google to “pull” into its system.

**This subcommittee recommends a “pull” strategy that incorporates BOTH standard and non-standard datasets.**

## Non-Standardized Datasets

A strategy to embrace non-standardized datasets favors rapid start-up, simply by immediately using local and state health departments’ common practice of publishing inspection results to the web. By identifying and collecting or “spidering” existing public-facing web sites, the project can expect early returns.To grow the database, CFP, NEHA, and other leaders need only advocate for best practices… specifically, adopting the FDA Model Food Code and publishing inspection results to the web.

Within these non-standard datasets, we can expect to realize many of the stated goals of the committee’s charge. That is, with these data, we can present the frequency of violations with keywords and date ranges in dashboards dedicated to this purpose. Within these data, we may curate and stream data to brand owners for their own analyses.

As the Model Food Code is adopted more widely, one can expect these data to naturally normalize over time. As the tools and techniques mature, more and more can be determined from the universe of data, not unlike election pundits and news organizations which increasingly scour and extract their conclusions from data-streams such as Twitter and others.

The costs and management under this approach are centralized. The primary request of our state and local partners is to continue along the course already set… to embrace the FDA Model Food Code and to publish their inspection results to the web.

**This subcommittee recommends a system capable of discovering, collecting, and indexing the immediately-available restaurant inspection results published to the web by health departments, even if these results do not meet a documented format or standard.**

## Standardized Datasets

The subcommittee would never dismiss the added value of standardized inspection datasets. After all, standardized data is part of the committee’s charge.

In fact, where a health department is able and willing to adhere to a published data schema and make that data file available in a "pull" configuration, the recommended system must routinely retrieve the updated dataset in favor of any corresponding web-published restaurant inspection pages.

We must recognize that data standards are adopted slowly. A data standardization project should begin with a version-controlled schema that captures the elements described by Annex 7 of the 2009 Model Food Code - Form 3a. Further, the first outreach promoting adoption of a uniform data standard should be to the commercial off-the-shelf software providers where integrating the standard into a commercial system provides for greater reach as that system is adopted.

This subcommittee recommends a published schema for standardized data which, if provided by the health department, is retrieved and processed instead of the public web site data.

### Data Stewardship

Data stewardship refers to the person or entity responsible for the data content and context. In practical terms, it means the entity who can correct errors and answer questions.

The subcommittee recommends that the health department that created the data remain the stewards of the data. Any data remediation should be brought to the attention of the originating agency, corrected in-place, and re-published.

## System Infrastructure

### Infrastructure Requirements

The following high-level requirements are addressed by this proposed infrastructure.

| Category | Requirement | Notes |
|---|---|---|
| Web Site Crawling Capacity | 18M web pages (inspections) per week[2] | Assuming:<br>- 3,000 health departments[3]<br>- Average 2,000 facilities per department<br>- Average 3 inspections per facility |
| Standardized File Retrieval | 200 files per week | Assuming 200 state/local agencies provide consolidated file per published standard |
| Storage Capacity | 2 Million Inspection Events Annually<br>200 Million Checklist Items | Based upon a 1M estimated national inventory of restaurants, assuming average two inspections annually. |

[2] Most will be duplicates of previously crawled pages. Very few will be new/updated data.

[3] See http://www.naccho.org/topics/infrastructure/profile/upload/LHD_Workforce-Final.pdf

| | | |
|---|---|---|
| | Annually<br>10 Years History Maintained[4] | Checklist Items describe a distinct inspection data-point (e.g., IN, OUT, NA, NO). |
| Users | 1,000 Registered Users<br>250,000 Anonymous Simultaneous Users (Maximum capacity - in cases of national interest) | Registered include those with privileged access for query/research/study.<br>Anonymous Users include non-privileged public-access, including consumers, owners/brand-holders, news media, etc. |
| Interfaces | Login / Account Management / Security Aspects<br>Website Registration, Web Crawler / Data Discovery / Data Collector<br>Public-Facing Web Site for Search/Results and Data Export | Login / Account Management refers to the creation and approval of accounts as well as lost passwords, changed profile, etc. |
| Documentati on | End User Documentation (Online)<br>Administrator Documentation<br>Standardized Data File Format Documentation | |

## Infrastructure Recommendations

### System Development

The subcommittee recommends that the system be developed by a professional entity selected on the basis of its demonstrated technical aptitude as well as a minimum of five years' involvement in and understanding of the food safety regulatory principles at national level.

### System Requirements

In coordination with the CFP, this committee and other stakeholders must establish a Software Requirements Specification (SRS). The SRS may become the basis of an invitation to to bid on the project.

### System Testing

The project shall include both functional and load testing by the vendor and a committee of beta testers. The project shall also include initial and routine security vulnerability assessments.

### System Acceptance

System acceptance shall be delegated to a committee of stakeholders and beta testers.

[4] Although this may exceed the data/record retention policies of a reporting agency, the system may be designed to anonymize data for inspections that are no longer subject to record retention.

### System Maintenance

The resulting system will require ongoing maintenance including end-user support, change control management/implementation, optimization, monitoring, security patches, bug fixes, etc.

### Change Control Process

As a component of System Maintenance, the implementing entity shall recommend and facilitate a Change Control Process. The Change Control Process is the means by which bug fixes and enhancements shall be captured, prioritized, and addressed.

## Estimated Costs

| Task | Estimated Hours | Estimated Costs |
|---:|---:|---:|
| Inception - Gather Requirements | 250 | $37,500.00 |
| Inception - Author Software Requirements Specification | 250 | $37,500.00 |
| Inception - Manage Vendor Selection | 200 | $30,000.00 |
| **Total** | | **$105,000.00** |
| | | |
| Implementation - Infrastructure / Data Center | 85 | $12,750.00 |
| Implementation - Coding | 500 | $75,000.00 |
| Implementation - Testing | 250 | $37,500.00 |
| Implementation - Documentation - Including Standard Data Format | 120 | $18,000.00 |
| Implementation - System Acceptance | 60 | $9,000.00 |
| Implementation - Web Site Plug-Ins | 500 | $75,000.00 |
| **Total** | | **$227,250.00** |
| | | |
| Annual Maintenance - Data Center | | $24,000.00 |
| Annual Maintenance - End User Support | | $30,000.00 |
| Annual Maintenance - Change Control Management | | $30,000.00 |
| Annual Maintenance - Security Updates | | $12,000.00 |
| **Total (Annual Costs)** | | **$96,000.00** |

## Estimated Cost to Health Departments to Participate

The costs below estimate the investment by an individual reporting agency (e.g., health department) to engage by contributing its data.

This may occur through Electronic Data Transfer (EDT), CSV Upload, Direct Entry, or "Screen Scraping."

Within each option, the reporting agency may engage their existing vendor (e.g., a national vendor) or their own internal IT resources.

| **Method of Participation** | **Description** |
|---|---|
| Non-Standardized Dataset from Public Posting of Restaurant inspections | If agency already publishing inspection results to the web and/or is already engaged in a project to do so, the cost is near $0.00.<br>If the agency has not yet committed to publishing restaurant inspection results to the web, the project may be $10,000 to $20,000 to do so using established vendor or internally developed system. |
| Standardized Dataset - Curated and provided by health department to be "pulled" by central system. | Established Vendors - $10,000<br>Internally Maintained - $10,000 |