Food for Thought: An Exploration into the Complexities of Analyzing Multi-structured Data in Today's Big Data World

Anjli Solsi, Anne Francomano, and Lisa Street

Abstract- The abundance of data in various formats in mainstream society has grown significantly over the past decade. The performance of research and statistical analysis of heterogeneous data sources is challenging without a single repository on which to base conclusions. In terms of nutritional education, there is more information regarding the nutritional content of various foods than ever before, as well as the ability to collect consumer-driven and biometric data. We explore the process and challenges of researching big data subjects that provide data across multiple structural formats, specifically nutritional data. This paper provides tools and practices that may support the analysis of data from heterogeneous data sources, all while focusing on consumer food nutrition data. We illustrate analysis on the individual structures as well as on the connection of both traditional structured and unstructured data using SAS and Tableau, one of many visualization tools. This paper details the methodology and research associated with illustrating the complexities of creating a single data repository to house approximately 250,000 different food items, their ingredient lists and nutritional contents, and culinary recipes from multiple data structures. This paper also focuses on achieving a single analysis across multiple repositories while aiming to gain valuable insights from the relationships, not just each individual data structure.

***Index Terms*—heterogeneous data structures, JSON, MongoDB, mySQL, nutrition, nutrition monitoring, proc sql, relational database, SAS, semi-structured data, structured data, Tableau, unstructured data**

I. INTRODUCTION

With the rapid growth of data generation and collection across all industries, the act of data processing has become increasingly important. Many big data subjects provide data across multiple structural formats; however, there does not appear to be a specific tool to accomplish consolidating the various structures into a single repository. This specific topic of nutrition stems from the increase in nutritional information available, as well as the consumers desire to know what is entering their bodies for a variety of health and lifestyle reasons.

The research aims to inform users on the complexities associated with the creation of a comprehensive consumer food and nutrition database from various data structures. Along with the challenges encountered with the creation of this single repository, this research focuses on the analysis of data from heterogeneous data sources. The research aims to provide readers with a better understanding of the mechanics and challenges of providing information from multiple forms of data (i.e. structured, semi-structured and unstructured data) for consolatory analysis.

Three different nutrition and culinary data sources were chosen as the focus of this research. The structured data comes from publicly distributed information in the Branded Food Products Database, which was created by the USDA. The data details nutritional information and serving sizes of different branded food products. The semi-structured data was taken from Fourmilab and contains the caloric data of various foods. The unstructured data consists of recipes manually entered from various cookbooks.

The structured data was imported as individual relations into a MySQL database and each table was joined. A final, cleaned dataset was imported into SAS for individual structure analysis. The semi-structured data was scraped from the website using RStudio to get the data into an XML format and then converted to a JSON file for analysis. For the unstructured data, hand typed recipes were entered into a text file and converted to JavaScript for use in MongoDB. Manual data review and querying allowed for insights to be drawn between the BFP data and hand typed recipes.

The different types of analyses performed, both single structured analysis and the approach of joining different structures, provided valuable insights into the challenges of analyzing the data and the data itself. Specific rankings based on data such as calorie counts and manufacturers from the BFPD are shown. The calorie counts across food groups is displayed from the semi-structured data. Although unable to simply combine the different data structures, manual manipulation allowed for greater insight by joining the BFPD data to the recipe data, which is visually demonstrated in Section VI.

Research Paper submitted on December 12, 2018.
Anjli Solsi (e-mail: asolsi@mail.smu.edu).
Anne Francomano (e-mail: afrancomano@mail.smu.edu)
Lisa Street (e-mail: lstreet@mail.smu.edu)
The authors are graduate students at SMU.

In this exploration into the analysis of heterogeneous data structures, specifically nutrition related data, the complexities of this approach emerge. Based on this research, there does not appear to be a single or prevalent tool for merging multiple data structures into a single repository. Certain analytical and visualization tools, such as SAS and Tableau, offer insights into individual analysis of each structure and across repositories when certain conditions are met. However, in order to gain more valuable and broader insights, significant manual manipulation is required.

This paper begins with discussing the overall approach taken to analyze the different data structures. There is then a discussion of various efforts and ongoing research involving large-scale nutrition related data. This transitions into identifying the sources and elaborating on the three sample data sets from the three different data structures (i.e. structured, semi-structured, and unstructured). The process and validation section details exactly what steps were taken using specific tools to accomplish the desired analyses. After going through the process, the analyses performed, and insights drawn on each data structure are detailed. Throughout this research, different challenges were encountered in dealing with different data structures, and those are expanded on after explaining the analyses. This paper ends with mentioning the ethical considerations that must be made when dealing with different types of nutritional data, as well as our team's takeaways from this research.

II. Research Methodology

The research aims to focus on identifying the challenges encountered in dealing with structured, semi-structured, and unstructured data. This relates to cleaning, downloading, and mining the data. Pros and cons of each step will be noted as well as details highlighting the overall issues found in setting up the system.

First, the USDA Branded Food Products Database (BFPD), which contains consumer product nutrition information, was located, along with its published documentation detailing how to handle missing data, available API information, and file formats [6,7]. To create the relational database, the Excel data files were downloaded from the BFPD. In creating the database, consumer product nutrient analysis began in a relational database format. Table and data evaluation with the ER model was created. The database view was also cleaned prior to upload into SAS. Then, the finalized data set was imported into SAS for analysis purposes. The SQL data view was uploaded, and the 'proc sql' command was used for data manipulation and review. Data analysis of BFPD data was performed in SAS and Tableau and graphics were included for visual representation of the data.

For the semi-structured data, a JSON file of calories in various foods was generated and uploaded to MongoDB. This was accomplished by scrubbing data from an online source, Fourmilab, using RStudio to get the data in XML format. In order to perform the analysis, the XML format was converted to a JSON file. The analysis of this data was performed in Tableau to create graphics that deliver the information in a clear and concise manner.

The unstructured data consists of manually entered recipes from cookbooks into a text file. In order to upload the data to MongoDB, the text file was converted to a JavaScript file. Similar to the semi-structured data, this individual analysis was performed in Tableau to present the recipes and define the necessary ingredients for each dish.

The final step of this research involved querying and cross-analysis of the structured and unstructured data. Tableau was chosen as the preferred tool, having the ability to connect multiple data sets of different structures. The recipe data set was joined with the BFPD data to gain insights into the ingredients used in recipes, as well as the composition of those specific ingredients. This approach allowed for comprehensive insights and visualizations.

III. Related Work

The potential use of nutritional information is only limited by the human imagination. There is a wide range of research being applied and related entrepreneurial endeavors being pursued where large scale nutrition-related data is concerned, from food safety traceability to custom nutrition plan development and nutrition habit monitoring. These activities utilize a combination of traditional structured databases and more recently available unstructured data, such as that gathered from wearable technologies or the spoken word.

While Marvin *et al.* address the potential for mobile technologies to benefit food-safety, the key system components and workflow that they present are directly applicable to any core system that will incorporate nutritional data. The typical big-data workflow stages they cite are: data collection, data storage and transferring, data analysis, and data visualization [1].

Firms such as IBM recognize the potential market and have filed patents for mobile applications that utilize sensors that will not only detect user biometric data but also utilize that data to suggest recipes based on a nutritional-level evaluation of the user [2]. Firms such as habit LLC (www.habit.com), in which the Campbell Soup company has invested, base their services on research described by van Ommen *et al.* as "a systems biology–based approach" which includes metabolic and genetic analysis in order to create personalized nutrition plans based on submitted bio-material [3].

Integrating natural language processing as a mode of tracking individual's consumption habits in a more natural manner is a promising alternative to other methods, such as self-reporting via questionnaires [4]. Research conducted by Hezarjaribi *et al.* has resulted in Speech2Health, "a voice based mobile nutrition monitoring system." [4] Once the speech analysis portion is complete, the system relies on core food nutrition data to create a custom nutrition profile [4].

There are also concerted efforts in the EU to partner researchers with consumers for long-term public funded research projects. For instance, Richfield's is working on establishing a research infrastructure, i.e. consumer data platform, that connects core nutrition data with consumer-generated data and formal research [2]. Zimmermann *et al.* describe how the platform "will make available data for researchers interested in studying food and health-related consumer behaviour. "[5].

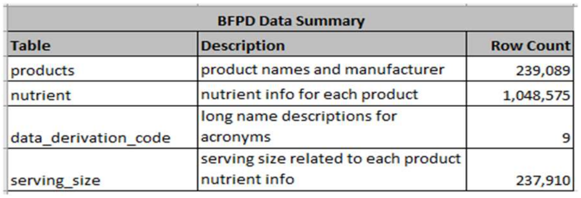
From the perspective of generally working with a mix of structured and unstructured data, Jaybal, *et al.* [12] propose a framework for analyzing data from heterogeneous data sources that incorporates an ontology-based approach to account for data relationships with respect to context, semantics and domain-specific vocabulary. While their framework utilizes transportation domain data from the Bangalore Metropolitan Transport Corporation (BMTC), their approach is conceptually applicable to other domains. Its data set is a representative example of heterogeneous data sources, consisting of schedules gleaned from webpages, csv files containing passenger survey feedback, freeform text from Twitter with passenger feedback and a relational database containing bus route Global Positional System (GPS) data.

IV. DATA

In today's big data world, the complexity of analyzing data across multiple structures is often daunting and difficult. Without the ability to control the structure of data, researchers are frequently left with large swaths of multi-structured data from which complicated and extensive analyses must be performed, and while there are many tools available to perform analysis in a singular fashion, the ability to perform analysis across heterogenous data structures in a unified form is limited due to a lack of automated resources available. So how does a person or organization achieve a more unified analysis across multiple data structures? This is the question posed within the following sections of this paper. To illustrate these difficulties, the subject of nutrition and nutritional education are discussed, including three sample data sets from three different data structures (i.e. structured, semi-structured and unstructured).

Nutritional research is an ever-evolving field of study, but modern nutritional science is still relatively young. It was not even until 1926 that the first vitamin was isolated and chemically defined, and even more recent that the role of nutrition and disease came into focus. "Research on the role of nutrition in complex non-communicable chronic diseases, such as cardiovascular disease, diabetes, obesity, and cancers, is even more recent, accelerating over the past two or three decades and especially after 2000." [8]. With greater interest into the study of modern nutrition, massive swaths of nutritional data are readily accessible and available for download, existing in both structured, semi-structured and unstructured forms.

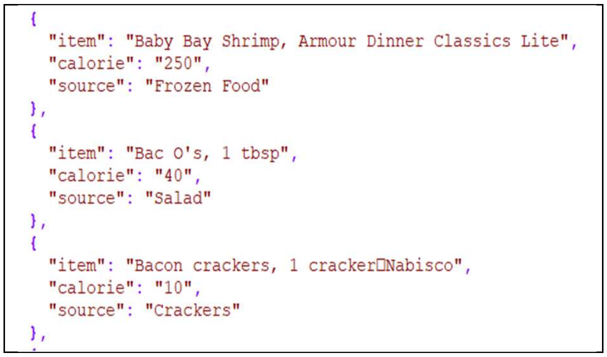
The first example of nutritional data comes from data distributed for public consumption in a structured format via the Branded Food Products Database (BFPD) [6]. This database was created by a Public-Private Partnership between staff members of the Nutrient Data Laboratory, Beltsville Human Nutrition Research Center, Agricultural Research Service and the US Department of Agriculture. Containing over 235,000 different branded food products, this database provides detailed nutritional figures and serving size details of those products. In addition to the product, nutrients and serving size data sets, the BFPD also provides details of each data set's file format as well as how to interpret missing data and zero values. The nutrients data set was also pre-standardized in terms of unit basis where these data were converted to a 100-unit basis, either gram (g) or milliliter (ml), depending on which was received from the data provider. Figure 4.1 provides a summary for each data table residing within the BFPD database.



BFPD Data Summary		
Table	**Description**	**Row Count**
products	product names and manufacturer	239,089
nutrient	nutrient info for each product	1,048,575
data_derivation_code	long name descriptions for acronyms	9
serving_size	serving size related to each product nutrient info	237,910

Fig. 4.1: Tables in the BFPD structured data source

The second example of nutritional data is provided in a semi-structured format via the Fourmilab website [9]. The caloric data from various foods displayed within this website were scraped into a JSON file format for consumption and analysis. Although more limited in its nutritional detail offering than the BFPD data set, the Fourmilab data also provides an extensive list of branded and non-branded food products and their caloric content. Figure 4.2 displays a sample of the data structure within the JSON file.



```
{
  "item": "Baby Bay Shrimp, Armour Dinner Classics Lite",
  "calorie": "250",
  "source": "Frozen Food"
},
{
  "item": "Bac O's, 1 tbsp",
  "calorie": "40",
  "source": "Salad"
},
{
  "item": "Bacon crackers, 1 cracker□Nabisco",
  "calorie": "10",
  "source": "Crackers"
},
```

Fig. 4.2: Sample semi-structured data format

Finally, the last example of nutritional data is in an unstructured format, via manually typed recipes from various cookbooks and websites. Without a pre-defined data model, the data captured within this data set encompass irregularities and ambiguities that may make it more difficult to initially extract meaning through simple analysis. For example, one recipe includes an ingredient of Old El Paso Picante Sauce. If a researcher were to query both ingredients and recipe origins, the term 'El Paso' may stand out in that it represents both a city in Texas and part of a branded ingredient name. Without the structure to define this attribute, an added level of complexity is apparent, as demonstrated in Figure 4.3 below.

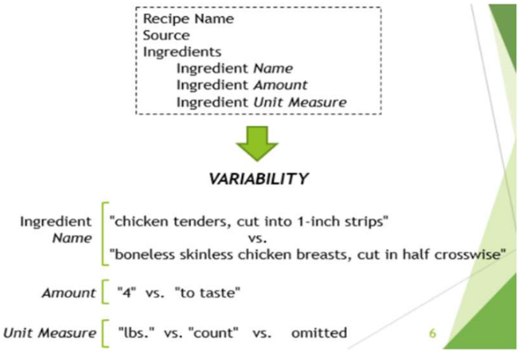

Recipe Name
Source
Ingredients
Ingredient *Name*
Ingredient *Amount*
Ingredient *Unit Measure*
VARIABILITY
Ingredient Name
"chicken tenders, cut into 1-inch strips"
vs.
"boneless skinless chicken breasts, cut in half crosswise"
Amount
"4" vs. "to taste"
Unit Measure
"lbs." vs. "count" vs. omitted
6

Fig. 4.3: Variability of unstructured data in culinary recipes

V. PROCESS AND VALIDATION

For the structured BPFD data, each unique data set (i.e. product, nutrients and serving size CSV file) is imported as individual relations into a MySQL database. Due to the structured nature of these files, each table is then joined together by a unique identifier ("NDBNumber") between the product, nutrients and serving size relations. From there, a master file of all the relation information is generated and exported as a new data set, to allow for import into our various analysis tools (i.e. Tableau, SAS).

The task of gathering the unstructured data into the JSON file format requires a bit more work. The data within the Fourmilab website is scraped via RStudio by taking each unique html table present within the html nodes of the Fourmilab webpage (of which there were thirty-five) and combining them into a single data frame. The data frame is then exported into an XML format via the open source RStudio XML library. At this point, in order to present the unstructured data in a format that is more widely accepted by our planned analysis tools, the XML is then converted to a JSON file format via a free online conversion tool [11].

For the fully unstructured data, various recipes found both online and in hard-copy cookbooks are hand typed into a text file to ensure no specific structure is obtained. The text-based recipe data is then converted to a JavaScript file for import into a MongoDB database and collection. Without any specific structure to the recipe data, the MongoDB database is an ideal repository for this data.

Since the structured and semi-structured data are both produced from online sources with details of the data presented within their source sites, validation of completeness and accuracy is performed through either simple queries or manual data review. For example, the structured BFPD data that is loaded to the MySQL database is queried by using the SQL 'count' function to ensure each row is present within each relation. Manual selects of each relation are also performed to ensure no data truncation has occurred. The semi-structured data which is scraped via RStudio and ultimately transformed into a JSON file format, is reviewed within RStudio for accuracy and completeness and then upon upload to SAS and Tableau, is simultaneously validated for completeness using functions like proc SQL and simple select statements. The unstructured data which is loaded into the MongoDB via a shell command on a JavaScript file, is also queried using the 'Count' and 'getCollectionInfos' functions available within MongoDB.

VI. ANALYSIS & RESULTS

Analysis of the three data sets occurs primarily within two tools, Tableau and SAS. Tableau is a desktop business intelligence and data visualization tool with a very intuitive UI, allowing for users with little to no coding experience the ability to freely navigate and create informative reporting. SAS is an integrated system of software solutions that enables users to not only retrieve, store and manipulate data, but also to perform complex statistical and predictive analysis as well as graphical reporting. Both SAS and Tableau are chosen for this project because of their ability to join and report on multiple data sources originating from different data structures. However, the ability to join and report on multiple data sources and across multiple data structures may require significant data massaging and perhaps even manipulation to reflect an accurate cross-analysis.

A singular analysis of each unique data set is often the best place to begin. Reviewing and analyzing data components within each data set will give actionable insights to individuals and organizations as to what is being represented. Using SAS for example, we are able to load the structured BFP data and run queries to find the top ten manufactured foods with the highest calorie counts. Figure 6.1 below provides a graphical view of these ten food products to quickly assess the caloric content of each food item.

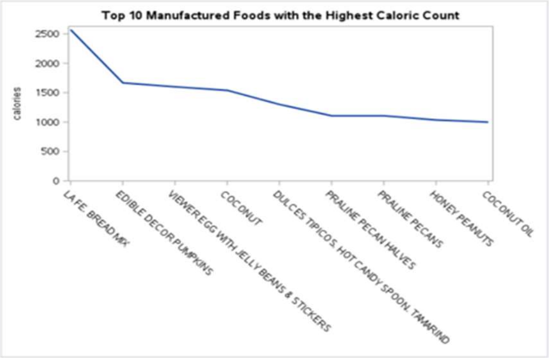


Fig. 6.1: SAS chart based on BFPD structured format

Our semi-structured data set also gives calorie counts per food item, and better yet provides food groupings so we can determine calorie content across an entire food group. Figure 6.2 below shows an analysis in Tableau of the Fourmilab semi-structured data across the Meat, Fish and Poultry foods groups, and provides an average calorie count across each group.

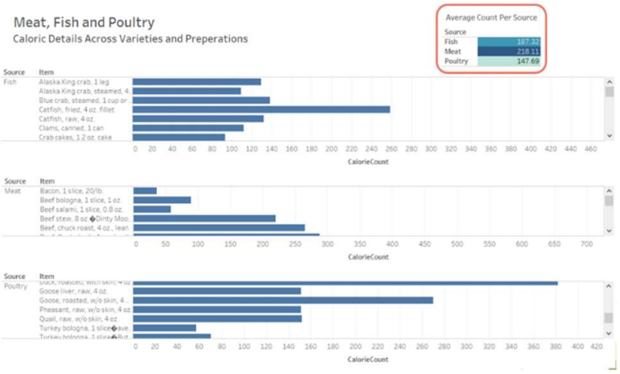


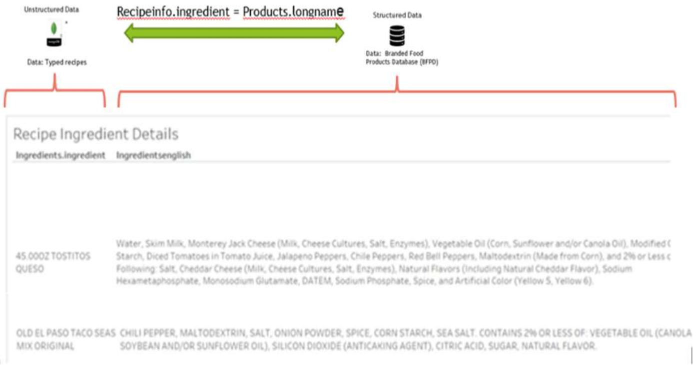
Fig. 6.2: Tableau analysis based on semi-structured format

However, single structured analysis of unique data sets can present a siloed approach to analysis, especially when the end goal is gathering large-scale insights. While all data originating from different data sources is not perfect in the sense that it will likely not join flawlessly together without some sort of manual intervention, the complexity of trying to join various data sets with different data structures is an even greater challenge.

Due to the abundance of data in various formats and residing across multiple platforms, the skillsets of Data Scientists and/or Data Engineers may be utilized to properly align, and re-structure unstructured data as needed. In an online article entitled 'There's no such thing as unstructured data' by Chuck Densinger and Mark Gonzales [10], restructuring unstructured data is noted to be a likely requirement for most organizations today, and details some worthwhile steps to take to provide better structure and meaning to unstructured data. These steps include:

1) Use data scientists and data engineers to uncover patterns which can be applied to give structure to the object(s) on which you plan to analyze
2) Transform and load the valuable unstructured data identified by the data scientists and engineers, into a structured format of data tables and attributes.
3) "Take a use-case driven approach to implement specific operational uses of unstructured data." [10]
4) Leverage purpose-build tools like RStudio to pre-process valuable data after patterns are recognized from use-cases.

By incorporating the steps mentioned above [10] into this analysis, we are able to gain better insights across our data sets. For example, the BFPD structured data contains food products that are used as ingredients in the manually typed unstructured recipe data. In order to allow for a join between these two data sets, manual manipulation of the recipe data set occurs by augmenting the original data with an additional column that explicitly ties the recipe ingredient name to the BFPD product. Figure 6.3 below demonstrates how in Tableau, the BFPD data joins to the massaged unstructured recipe data set, allowing for greater detail of the ingredients within the ingredients; that is, the ingredients that makeup an ingredient listed in a recipe. We see in this figure that the recipe ingredient of Tostito's Queso from the recipe data, also exists and has detail in the BFPD structured data, which provides greater insights to this ingredient's makeup.


Unstructured Data
Data: Typed recipes
Recipeinfo.ingredient = Products.longname
Structured Data
Data: Branded Food Products Database (BFPD)

Recipe Ingredient Details

Ingredients.ingredient	Ingredientsenglish
45.00OZ TOSTITOS QUESO	Water, Skim Milk, Monterey Jack Cheese (Milk, Cheese Cultures, Salt, Enzymes), Vegetable Oil (Corn, Sunflower and/or Canola Oil), Modified (Starch, Diced Tomatoes in Tomato Juice, Jalapeno Peppers, Chile Peppers, Red Bell Peppers, Maltodextrin (Made from Corn), and 2% or Less c Following: Salt, Cheddar Cheese (Milk, Cheese Cultures, Salt, Enzymes), Natural Flavors (Including Natural Cheddar Flavor), Sodium Hexametaphosphate, Monosodium Glutamate, DATEM, Sodium Phosphate, Spice, and Artificial Color (Yellow 5, Yellow 6).
OLD EL PASO TACO SEAS MIX ORIGINAL	CHILI PEPPER, MALTODEXTRIN, SALT, ONION POWDER, SPICE, CORN STARCH, SEA SALT. CONTAINS 2% OR LESS OF: VEGETABLE OIL (CANOLA SOYBEAN AND/OR SUNFLOWER OIL), SILICON DIOXIDE (ANTICAKING AGENT), CITRIC ACID, SUGAR, NATURAL FLAVOR.

Fig. 6.3: Tableau connects structured and unstructured data

VII. CHALLENGES

An initial challenge encountered during our research was determining whether to try to merge the structured, semi-structured and unstructured data all into a single 'master' database. The alternative was to instead preserve these three separate, disparate data repositories and utilize Business Intelligence (BI) reporting tools, such as Tableau, to create dashboards based on joining the tables within the various databases.

It became clear that the ultimate repository structure is not the sole critical factor for facilitating analysis. Effective data mapping is a key factor as well. Regardless of whether the data has been merged or not, it is the nature of the data components themselves and the relationships of the data across repositories on which data mapping depends. A longstanding obstacle facing any data integration effort is "how to universally map information from heterogeneous data sources" [13]. Being able to join across heterogeneous data structures relies on successful data mapping. The manner in which the underlying data were captured may also contribute to potential data variability. The process to address data variability may involve a large degree of manual data manipulation in order to establish consistency across data sets, such as standard naming convention.

There is currently a dearth of software tools offering automated analysis of unstructured data. Such tools would resolve the complexities inherent in analyzing data from heterogeneous data structures. The need to address this problem is evident by the fact that commercial organizations have filed related patent applications, such as for approaches that incorporate machine learning-based extraction and classification methods [16].

In addition, even though some data structures are easier to query on and consume than others, there may be significant time spent up front by developers on initial tool configuration. This is not an uncommon occurrence when working with software in general, yet it can be unpredictable. For example, in this project, establishing the Tableau connection to MongoDB was unexpectedly time consuming.

VIII. ETHICS

In any research-based project, there are always ethical considerations with respect to the data origin, collection process, conversion activity, and implementation of results.

Regardless of the number of data sources being utilized, the validity of each source must be evaluated. When nutrition data is provided directly from a manufacturer, such as via physical packaging labels in accordance with legal regulations, consumers trust that the data is accurate. Similarly, if a system is being created to support a new product offering, such as a personalized nutrition advisory application, it is important that users can be assured that they are relying on trustworthy, consistent sources.

There may be ethical implications associated with the collection methods of consumer and commercial related data. Data collection methods must aim to be unbiased. Once data resides in a repository, subsequent data manipulation by other parties may be necessary in order to ensure consistency for comparison and reporting purposes, such as standardizing the units of measure. The structured BFPD utilized in this research relies on "a number of food industry data providers" such as Label Insight and 1WorldSync to whom companies voluntarily submit data by means of the Global Data Synchronization Network. While these food industry data providers are responsible for providing the core, underlying data being shared by the USDA [7], the data received by the USDA must first be standardized prior to being reporting. Any necessary reformatting and standardizations are performed together by the USDA and the University of Maryland's Joint Institute for Food Safety and Applied Nutrition. The USDA documentation details the conversion approach and notes areas where any additional resulting variability may be introduced. The sharing of the data origin and conversion details is important to maintain ethical transparency with users.

When data are pulled in from heterogeneous data repositories for comparison purposes, additional data cleaning may be necessary. For example, while each repository may hold data for a manufacturer's products, each repository may employ its own product naming conventions. Also, one repository may hold data for a product not tracked in the other repository. Options for aligning data entries for comparison may include setting up a mapping between product names or even modifying the local repositories so that the names match. Data cleaning has the potential to introduce errors and unintentional bias in product name mapping.

As is the case across a variety of online data topics, nutrition-related data is available in multiple forms across multiple platforms. Therefore, it is not always feasible to recognize what might be an individual's opinion or a lack of scientifically established protocols in their research. The Fourmilab semi-structured caloric data [9], for instance, does not cite its source (e.g. from set reference material or personal calculations) and could contain misleading information.

The generation and use of results for personalized nutrition services, such as those provided by habit LLC that include genetic-based analysis [3], present an additional set of ethical concerns [14], particularly where protection of genetic data is concerned. The veracity of the personalized, analytical results is one concern, along with how well the genetic material, its associated data and analytical results are safeguarded. It is imperative that consumers of nutritional information analysis remain cognizant of the fact that the science and study of modern nutrition is a work in progress. With the legal framework that applies to genetic material typically lagging behind the latest technological developments [15], solution providers and consumers should proceed with caution and understand current legal regulations and possible limitations.

IX. CONCLUSION

The momentum of data generation and acquisition shows no sign of slowing down anytime soon. While nutrition and culinary recipe data is used in this research for illustration purposes, it is merely one of countless domains where shared issues emerge pertaining to data analysis from heterogeneous data structures.

Currently there does not appear to be a single tool or toolset in wide use for merging data from disparate data structures into a single repository. There are disadvantages to merging the data, such as data loss. Instead, the goal could be to establish valid relationships between the data across these mixed sources within the domain-specific context, such as that of nutrition, transportation, etc. However, there are also limited resources for the automation of relating data across multiple heterogeneous sources. Without remedies, silo-centric analysis will likely perpetuate and impede broad-scale analytical efforts.

Visualization tools, such as Tableau, are effective at tapping into a variety of data structures individually and across repositories if the necessary data consistency is in place, such as naming conventions. However, the pre-analysis data preparation tasks remain daunting. Typically, significant data cleansing and data manipulation are required in order to gain broad insights across multiple data structures. While this currently manually intensive work may be worth the effort if beneficial insights are gleaned in a timely manner, it is not sustainable over the long term as additional types of data become available and data stores continue to grow.

As society's dependence on large-scale data continues to expand and evolve, the methods we employ to process and interpret this data become even more important. These methods are truly a work in progress. Current efforts to address determining the relationships between data, and resulting interpretation such as classification, include a proposed framework that incorporates machine learning in conjunction with domain-specific vocabulary [12]. An aspect to consider for future work is exploring the role of machine learning in determining which pieces of data, within a literal sea of data, could be deemed useful. Usefulness tends to be subjective in nature. One way to define usefulness is within the context of one or more problems at hand that need to be solved. Perhaps machine learning could be put to work for this complex purpose.

REFERENCES

[1] Hans J. P. Marvin, Esmée M. Janssen, Yamine Bouzembrak, Peter J. M. Hendriksen & Martijn Staats (2017). Big Data in Food Safety: An Overview, Critical Reviews in Food Science and Nutrition, 57:11, 2286-2295 https://doi.org/10.1080/10408398.2016.1257481

[2] Timothy J. Chainer, Gerard McVicker, Pritish R. Parida; IBM, Patent application for Personal Food Database (PFD) (2016). https://patents.google.com/patent/US20180061270A1/en

[3] B. van Ommen, T. van den Broek, I. de Hoogh, M. van Erk, E. van Someren, T. Rouhani-Rankouhi, K. Hogenelst, W. Pasman, A. Boorsma, and S. Wopereis are with TNO (The Netherlands Organization for Applied Scientific Research), Zeist, the Netherlands. J.C. Anthony is with Habit LLC, Oakland, California, USA. Systems Biology of Personalized Nutrition. Nutrition Reviews®. Vol. 75(8):579–599. (2017). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5914356/pdf/nux029.pdf

[4] Dean Niloofar Hezarjaribi, Student Member, IEEE, Sepideh Mazrouee, Student Member, IEEE, Hassan Ghasemzadeh, Senior Member, IEEE. Speech2Health: A Mobile Framework for Monitoring Dietary Composition from Spoken Data. DOI 10.1109/JBHI.2017.2709333, IEEE Journal of Biomedical and Health Informatics (2016). http://epsl.eecs.wsu.edu/wp-content/uploads/2015/03/Niloofar_JBHI.pdf

[5] Karin L. Zimmermann, Muriel Verain, Marc-Jeroen Bogaardt, Anouk Geelen, Paul Finglas, Pieter van 't Veer, Monique Raats, Bent Egberg Mikkelsen and Krijn Poppe. Linked Data Sharing to Foster Consumer Based Science Enabled by Richfields: A Research Infrastructure on Consumer Health and Food. *Measuring Behavior 2016, 10th International Conference on Methods and Techniques in Behavioral Research (25-27 May 2016).* https://www.richfields.eu/wp-content/uploads/2017/02/Measuring-behavior-2016-paper-Linked-Data-Sharing-to-Foster-Consumer-Based-Science-Enabled-by-Richfields.pdf

[6] USDA Branded Food Products Database (BFPD) data source (as of July 13, 2018) https://ndb.nal.usda.gov/ndb/

[7] USDA Branded Food Products Database (BFPD) Documentation and Download User Guide (August 2018) https://www.ars.usda.gov/ARSUserFiles/80400525/Data/BFPDB/BFPD_Doc.pdf

[8] Dariush Mozaffarian and colleagues: *History of modern nutrition science – implications for current research, dietary guidelines, and food policy* (22 October 2018). https://www.bmj.com/content/bmj/361/bmj.k2392.full.pdf

[9] John Walker: Fourmilab, *Calories in Various Foods (2005).* https://www.fourmilab.ch/hackdiet/e4/foodcalories.html

[10] Chuck Densinger and Mark Gonzales: *There's no such thing as unstructured data.* Analytics Magazine (September/October 2016). http://analytics-magazine.org/theres-no-such-thing-as-unstructured-data/

[11] Domenico Briganti: *utilities-online.info (last modified December 6, 2018).* http://www.utilities-online.info/xmltojson

[12] Yogalakshmi Jaybal, Chandrashekar Ramanathan, S. Rajagopalan: HDSAnalytics: A data analytics framework for heterogeneous data sources. (2018) *ACM International Conference Proceeding Series*, , pp. 11-19.

[13] Arkopaul Sarkar, Dusan Sormaz: Multi-agent System for Cloud Manufacturing Process Planning. 28th International Conference on Flexible Automation and Intelligent Manufacturing (FAIM2018), June 11-14, 2018, Columbus, OH, USA

[14] Kohlmeier, Martin & De Caterina, Raffaele & Ferguson, Lynnette & Görman, Ulf & Allayee, Hooman & Prasad, Chandan & X. Kang, Jing & Nicoletti, Carolina & Alfredo, Martinez. (2016). Guide and Position of the International Society of Nutrigenetics/Nutrigenomics on Personalized Nutrition: Part 2 - Ethics, Challenges and Endeavors of Precision Nutrition. Journal of Nutrigenetics and Nutrigenomics. 9. 28-46. 10.1159/000446347.

[15] Vivek Wadhwa. Laws and Ethics Can't Keep Pace with Technology. MIT Technology Review. April 15, 2014. https://www.technologyreview.com/s/526401/laws-and-ethics-cant-keep-pace-with-technology/

[16] Felix Mueller; Swiss Reinsurance Company LTD., Zurich (CH), Patent application for Data Extraction Engine for Structured, Semi-structured and Unstructured Data with Automated Labeling and Classification of Data Patterns or Data Elements Therein, and Corresponding Method Thereof, (2016).