VacSeen: A Linked Data-Based Information Architecture to Track Vaccines Using Barcode Scan Authentication

Partha S Bhattacharjee[1⋆], Monika Solanki[2], Rahul Bhattacharyya[1], Isaac Ehrenberg[1], and Sanjay Sarma[1]

[1] Auto ID Labs, Massachusetts Institute of Technology, Cambridge, United States
[2] Department of Computer Science, University of Oxford, Oxford, United Kingdom
parthasb@mit.edu, monika.solanki@cs.ox.ac.uk,
{rahul_b,yitzi,sesarma}@mit.edu

Abstract. Renewed global efforts to deploy Automatic Identification and Data Capture (AIDC) technologies, such as barcodes, on vaccine packaging in developing countries are currently underway. An opportunity to evaluate Linked Data technologies for generating an ecosystem of data connectedness and interoperability in the vaccine supply chain presents itself. We discuss the VacSeen project, a Linked Data-based information system to track vaccines through visualization and authentication of barcode scans on vaccine packaging using mobile phones. The project is aimed at enabling endeavors such as logistical planning and integration with health information systems, demand forecasting, anti- counterfeiting and diversion measures, and post-marketing surveillance by pharmaceutical companies, supply chain contractors, and public health agencies. By forming an abstraction layer over siloed data while necessitating minimal modification of existing architecture, VacSeen can help minimize the technical, operational, and political friction often associated with fostering data interoperability. We discuss VacSeen's software architecture and present sample data analytics that highlight VacSeen's ability to facilitate the interoperability of diverse and non-standardized data sources. Limitations of the current framework and areas of future exploration and expansion are also discussed.

Keywords: Vaccine, Supply Chain, Event Modeling, EEM, LOD Cloud, Barcode, RDBMS2RDF

1 Background and Related Work

Vaccines have been globally recognized as a critical public health intervention [1]. Though vaccine access has improved globally, immunization rates in developing world are often sub-optimal at around 80% [1]. This subdued coverage is, in part, due to the near-absent visibility into the movement of vaccines in the supply chain. The lack of information renders demand forecasting difficult and limits

⋆ The authors would like to thank Richard Cyganiak for inputs on RDB-RDF conversion.

[1] http://data.unicef.org/child-health/immunization

the ability to effectively resolve issues associated with counterfeiting and product diversion. With the total vaccine consumption per child set to increase by up to 143% with new vaccination schedules [11], there is a dire need for technology and policy frameworks for improved track and trace visibility.

Several countries have issued directives to vaccine manufacturers to include barcodes on vaccine packaging [2,3] [4]. However, the adoption of such technologies has been historically lackluster in the developing countries because of issues such as absence or high cost of supporting infrastructure, need for skilled labor for operation, and lack of stakeholder engagement [personal communication, 2014]. With the advent of personal computing devices and improvements in wireless communication networks, the opportunity to use these AIDC technologies presents itself again. The Vaccine Packaging and Presentation Advisory Group (VPPAG), a joint effort by the major stakeholders in vaccine access, has launched a multi-stakeholder project in Tanzania to implement barcodes on vaccine packaging representing a renewed global effort to promote AIDC technologies in the vaccine supply chain [10]. For this project to succeed, demonstration of the additional value generated by deployment of AIDC technologies is critical. The first step to doing so is to provide vaccine consumption data — especially for the last mile of the supply chain. We discuss VacSeen, a Linked Data-based information system that attempts to provide such visibility. VacSeen takes into account the challenges posed by a global supply chain such as disparities in communication technology, non-standardized data storage, and the need for interoperability with electronic healthcare systems.

Numerous commercial entities are presently engaged in developing technology-based solutions for tracking vaccines in developing countries [4,5,6]. Despite improved information flow, issues about interoperability and at-scale last mile tracking continue to persist. Several studies have reported the effectiveness of using barcodes in tracking vaccine consumption [5], [2]. However, we are not aware of any studies that have leveraged Linked Data technologies for such applications, or adopted the approach of scan validation.

We demonstrate the utility of VacSeen in authenticating barcode scans and generating attendant rich contextual information. We geolocate and classify the scans based on whether or not they were recorded by authorized personnel using authenticated devices. We learned from vaccine manufacturers that such granular visibility into vaccine movement in developing countries does not presently exist. In this publication, we mimic data from multiple sources—a field-based mobile application, a supply chain database with information about barcodes, and a healthcare provider database with details about personnel and devices

[2] http://www.sidley.com/news/major-vaccine-standards-change-in-china-01-24-2011
[3] http://www.securingindustry.com/turkey-sets-short-timeframe-for-pedigree-system/s15/a29/
[4] http://vaxtrac.com/
[5] http://www.logistimo.com/products
[6] http://www.path.org/vaccineresources/supply-chain-and-logistics-systems.php

scanning the barcodes—to demonstrate that Linked Data technologies can be effectively used to render data from multiple sources interoperable. In addition, we leverage the Linked Open Data (LOD) cloud to enrich the dataset by generating biomedical factsheets about the vaccines as well as identifying nearest airport to a selected scan location. Such applications demonstrate the utility of Linked Data in activities such as knowledge management, logistical planning, and product recall.

Section 2 presents an overview of the VacSeen system architecture and presents each software component in detail. Section 3 presents proof-of-concept results obtained from VacSeen such as aggregate statistics of vaccine consumption and geo locations of healthcare workers handling vaccine packaging. Finally, Section 4 summarizes the key findings of our study and scope for future work.

2 Architecture

The presence of barcodes on primary and secondary packaging presents the opportunity of using an EPCIS v1.1[7](Electronic Product Code Information Services) event as a proxy for a vaccine transaction event. EPCIS is a standardized event oriented specifications prescribed by GS1, a standards organization,[8] for enabling traceability. Here a transaction event can either be traversal through the supply chain or the administration of the vaccine. By including barcode scanning as a required step in standard operating procedure, a record of every vaccine receipt event can be stored.

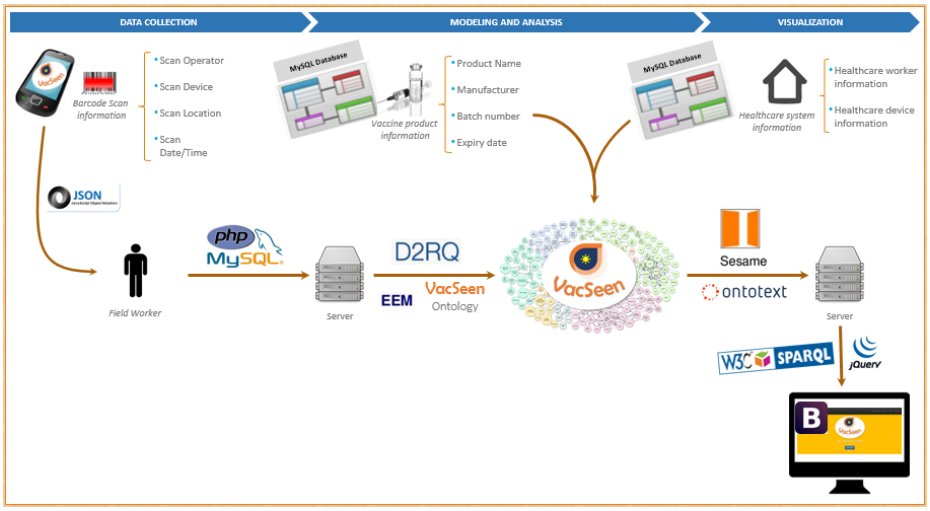


Fig. 1: The VacSeen project - System Architecture

We designed VacSeen's architecture using scalable open source tools. VacSeen comprises an Android application for barcode scanning, a server that hosts mul-

[7] http://www.gs1.org/gsmp/kc/epcglobal/epcis
[8] http://www.gs1.org/

tiple relational databases, an ontology for mapping the conversion of relational data into Linked Data, a triple store for storing the converted Linked Data, and a web-based visualization platform (Fig 1).

2.1 VacSeen Mobile Application

Our focus while developing the Android application was testing a Minimum Viable Product in the field that we can add features to at subsequent stages of development. As a result, the role of the VacSeen application was confined to that of a generator for scan data.

The worker is expected to scan the barcode on vaccine's package as an identifier of a transaction event. We integrated the widely used Zxing barcode library[9] into the application to facilitate barcode scanning. The app collects the following information that serve as components of the business rules for scan authentication:

- Content and format of the barcode scanned.
- Worker's phone number that serves as operator ID.
- The device's International Mobile Equipment Identity (IMEI) number that serves as device ID.
- Spatial and temporal data.

Despite the availability of several other device identifiers, we chose the IMEI number because of its relative ubiquity and uniformity. The capture of IMEI number can give rise to concern among users as the application requests access to call records during installation. However, we assume that the users of the application are authorized personnel using government- or employer- issued devices as is often the case with immunization projects. As a result, we circumvent concerns about privacy breach that would otherwise typically arise.

2.2 Data storage

For storing and hosting the data, we used WAMP Server (Apache, PHP, MySQL on Windows)[10]. The data from VacSeen mobile application was stored in a MySQL database hosted on an Apache server with PHP as the server side scripting language. A JSON parser in the mobile application was used for transmitting the data which was then written into the database by a PHP script. We chose WAMP Server to mimic the use of MySQL on Windows computers in the barcode project in Tanzania.

In addition to the database for storing inputs from the mobile application, we created another database of authorized operators (scanning personnel) and devices (mobile phones) to mimic those used by the healthcare authorities.

[9] `https://github.com/zxing/zxing`
[10] `http://www.wampserver.com/en/`

2.3 RDB-to-RDF translation

With numerous options available for Relational Database (RDB)-to-Resource Description Framework (RDF) translation [6], we employed the D2RQ Platform[11] for our project, despite R2RML being the W3C recommended standard, for two reasons. First, D2RQ bears deeper logical similarity to RDF compared to R2RML which is closer to relational databases. Unlike D2RQ, the complexity of mapping using R2RML is largely centered on querying the relational database using *rr:sqlQuery*. The second reason is the the ability of D2RQ to support R2RML, particularly when dumping relational databases as RDF. Using D2RQ enables us to not only leverage an extensively used publicly available stable tool but also be compliant with standards in the future.

We used the D2RQ mapping tool and the dump generator for the project. To customize the mapping produced by the *generate-mapping* script, we used resources from multiple well-known vocabularies such as *dbpedia-owl*, *foaf*, and *eem* in addition to a self-created one named *VacSeen1*[12]. The mapping files are publicly available[13].

We enforced integrity constraints on the data during translation by specifying datatypes in the mapping files. Additionally, we applied filters in the SPARQL queries to disregard records with incorrectly captured data fields.

2.4 Conceptualizing Domain Knowledge as Ontologies

For creating the VacSeen1 ontology, we reused sections of the EPCIS Event Model ontology (EEM)[14] and the Vaccine Ontology (VO)[15]. The incorporation of concepts from the two ontologies enabled us to seamlessly bridge logistical and biological information for our future applications. We generated persistent uniform resource identifiers (URIs) for the ontology elements and are currently working on making them de-referenceable. We used a light-weight ontology with just enough formalization to enable detailed querying. As the datasets attain greater complexity, it will be necessary to incorporate a higher degree of semantics within the ontology. However, we will position most of the complexity in our queries instead of the ontology in order to control for reasoning errors.

EEM is an OWL 2 DL ontology for modelling EPCIS events. EEM conceptualises various primitives of an EPCIS event that need to be asserted for the purposes of traceability in supply chains. Development of EEM was informed by a thorough review of the EPCIS specification and extensive discussions with trading partners implementing the specification. The modelling decisions [7] behind the conceptual entities in EEM highlight the EPCIS abstractions included in the ontology. In [8] a mapping between EEM and PROV-O[16], the vocabulary

[11] `http://d2rq.org/`
[12] `http://web.mit.edu/parthasb/Public/Vacseen1_ontology.owl`
[13] `http://web.mit.edu/parthasb/Public/Vacseen_Customized_Mapping.ttl`
[14] `http://purl.org/eem\#`
[15] `http://www.violinet.org/vaccineontology/`
[16] `http://www.w3.org/ns/prov-o`

for representing provenance of Web resources has been defined. The event history can be interrogated using PROV-O for recovering provenance information associated with the events.

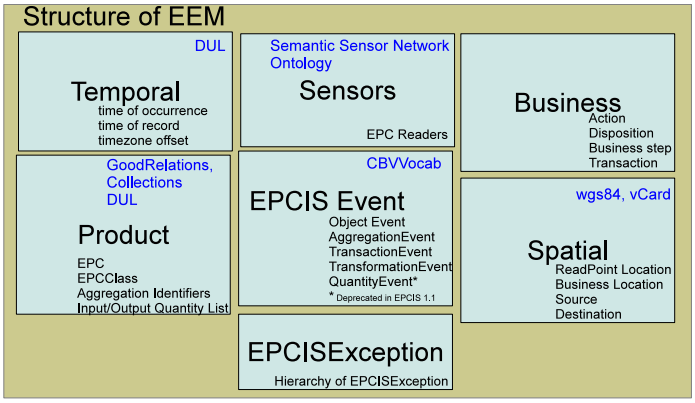


Fig. 2: Structure of EEM and its alignment with external ontologies (noted in blue coloured text)

The EEM ontology structure and its alignment with various external ontologies is illustrated in Figure 2. The ontology is composed of modules that define various perspectives on EPCIS. The *Temporal* module captures timing properties associated with an EPCIS event. It is aligned with temporal properties in DOLCE+DnS Ultralite (DUL)[17]. Entities defining the EPC, aggregation of EPCs and quantity lists for transformation events are part of the *Product* module. The GoodRelations[18] ontology is exploited here for capturing concepts such as an `Individual` Product or a lot (collection) of items, `SomeItems` of a single type. Information about the business context associated with an EPCIS event is encoded using the entities and relationships defined in the *Business* module. RFID readers and sensors are defined in the *Sensor* module. The definitions here are aligned with the SSN[19] ontology. The *EPCISException* module incorporates the hierarchy of the most commonly observed exceptions [9] occurring in EPCIS governing supply chains.

Since EEM offers an extensive model for EPCIS events, mapping the elements of the relational databases to classes and properties of the ontology was our default approach. However, we created additional properties and classes as needed. For instance, the location coordinates from the VacSeen application are stored in the *scan_event* table of the *vacseen_connect* database as *scanLat* and *scanLong* columns. To map the latitude and longitude coordinates, we created the *vacseen1:latitudeOfBarcodeScanEvent* and *vacseen1:longitudeOfBarcodeScanEvent*

[17] `http://ontologydesignpatterns.org/ont/dul/DUL.owl`
[18] `http://purl.org/goodrelations/v1`
[19] `http://purl.oclc.org/NET/ssnx/ssn`

properties in the VacSeen1 ontology. Other customized properties include *scanID* and *operatorID*.

The scanID is a 20-digit composite unqiue identifier of a scan event generated from scan attributes to facilitate an intuitive understanding about it. The information encoded in the ID can be useful in implementing access controls, protecting personnel privacy, limiting data exchange volumes, and partially offseting the ambiguity arising from multiple scans of unserialized GTINs.

2.5 Repository

We chose GraphDB-Lite[20] semantic repository implemented on Sesame for our project as it offers *owl* based reasoning and is free, stable, scalable, and easy to use because of an intuitive administrative interface that comes with the distribution. We used Apache Tomcat 8.0 servelet container as per the recommended installation settings on a personal computer.

We loaded data from the 3 databases —vaccine data, healthcare system data, and barcode scan data—(c.f Fig 1) into a single triple store. Additionally, we selectively incorporated relevant data for our analyses from the LOD cloud to circumvent the issue of sporadic unavailability of public SPARQL endpoints. By having the data in a single store, we excluded the need for more complex federated SPARQL queries that tend to be resource and time intensive. As the volume of data scales, the triple store data storage and access architecture will, of course, require re-design but we do not explore that in this paper.

2.6 Web Interface

The web interface of the project is publicly available[21] and was built using the Bootstrap framework. The interface presently enables querying a static dataset of EPCIS events to display 3 levels of scan authentication using Google Maps, attendant analyses, and Linked Data applications.

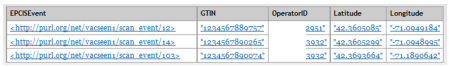
The marker data is generated by querying the Sesame triple store using the SPARQL query language over jQuery. In the absence of automated formatting of SPARQL queries in JavaScript, we formatted the queries through string concatenation.

3 Application

In this section we provide contextual information about vaccine scans, aggregate statistics, and description of two LOD-based applications. For the purposes of this study, we make use of simulated data where 22 researchers and volunteers in United States and India were asked to scan barcode labels (vaccines or otherwise) randomly over a period of 28 days.

The users generated 217 scan events for analysis of which 20 events from 3 users did not capture the location coordinates of the scan and returned values of

[20] `http://ontotext.com/products/ontotext-graphdb/graphdb-lite/`
[21] `http://parthasb.scripts.mit.edu/`



(a) SPARQL Output - Intermediate Validation Query

(b) SPARQL Output - Advanced Validation Query

(c) MySQL output - Eligible scans for Intermediate Validation

Fig. 3: SPARQL and MySQL Output for Validation

'0.0' for latitude and longitude. This is possibly a result of the lag experienced by devices at times to report their location coordinates. Subsequent versions of the mobile application will look to eliminate this problem by appropriate data buffering techniques.

3.1 Vaccine scan authentication

Several entities such as shipping agencies, inventory stock controllers, and healthcare workers, are engaged in vaccine logistics; all of whom can use VacSeen to scan barcodes during operations. It is therefore important to provide additional context for each barcode scan so as to get a deeper understanding about vaccine handling operations in the field.

To illustrate this, we randomly designated two authorized operators (IDs ending in '3932' and '2951') and an authorized device(ID ending in '1014' belonging to the first of the aforementioned operators. A scan by operator '3932' using device '1014', for example, can be used to indicate a vaccine administration scan by a healthcare worker while a scan by '2951' might indicate an inventory status scan by a non-healthcare worker.

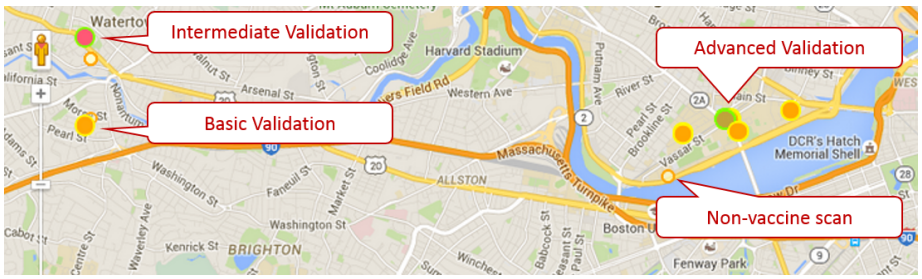


Fig. 4: Visualization of Validated Scans

Following are some authentication visibility statistics tested in the study:

– All Scans: Displays all of the 197 barcode scans (vaccine label or otherwise) by any operator and device using the VacSeen software.

- Basic Validation: Of the 197 scans, identifies the 37 scans that entailed scanning of a vaccine GTIN present in the supply chain database.
- Intermediate Validation: Distinguishes the 3 scans that were undertaken by the two authorized operators on vaccine GTINs (Fig 3a) as can be verified from the MySQL database (Fig 3c).
- Advanced Validation: Adds to the Intermediate Validation by distinguishing the 2 scans that were done using the sole authorized device. As the list of authorized devices only has one device registered to operator '3932', one of the scans registered in Intermediate Validation does not qualify as an Advanced Validation (Fig 3b). The differences among the validation levels are illustrated in Fig 4.

3.2 Operator and device scan statistics

In addition to barcode scan validation, we undertook preliminary analyses of the data to assess operator performance by measuring scans by devices and visualizing overall temporal trends in scans. Visualization of aggregate statistics was useful in determining individual and collective user activity which can be considered representative of analysis of personnel efficiency during enterprise applications.

3.3 Linkage to LOD cloud

To demonstrate the benefit of linkage to the LOD cloud, we generate biomedical factsheets about the vaccines comprising information such as type, route of administration, and Medline and Anatomical Therapeutic Chemical (ATC) numbers from DBpedia [3]. We use the *owl:sameAs* property to equate resources in the native dataset to those in DBpedia. We also developed an application to identify the nearest airport to a scan location to assist in endeavors such as logistical planning and product recall. For identifying the nearest airport, we use the *SPARQL SERVICE* feature and the *omgeo:nearby* property. This feature is a representative approach to geolocating other entities of interest corresponding to the scan sites. For instance, mashing scan density data with location of healthcare centers, hospitals, and warehouses can help generate rich insights about product movement and consumption. Such applications demonstrate that Linked Data technology can leverage the burgeoning open and structured data on the web more easily and seamlessly relative to relational database systems.

4 Conclusion and Future Work

In this paper, we report the development of VacSeen, a Linked Data-based information architecture for authenticating barcode scans on vaccine packages using mobile phones. We demonstrate how the software framework can be used to enable vaccine scan authentication visibility in vaccine transportation and administration logistics. For the purposes of this study, we make use of simulated data but the evolution of the project will be anchored to a pilot study.

We will extend the mobile application to facilitate two-way communication

so that the user can receive information compliant with access control mechanisms. Additionally, we will explore the possibility of capturing the data from the mobile device to a triple store directly.

With respect to data storage and querying, we will migrate to a cloud-based solution with due consideration to scale and security. We will continue to build applications leveraging the LOD cloud, such as generating comprehensive individual factsheet for every scan. Such a factsheet will draw from both native and external datasets. To manifest such applications, we are currently engaged in RDFlizing open datasets of interest. Other workstreams under consideration are using inferencing to classify data from external sources, and extending the VacSeen1 ontology to answer more complex questions in a broader set of use cases. Going forward, we will extend VacSeen in a bid to create an ecosystem of Linked Data with low adoption barriers to help address complex issues associated with transportation of healthcare goods and services to resource-constrained settings.

References

1. F. Andre, R. Booy, H. Bock, J. Clemens, S. Datta, T. John, B. Lee, S. Lolekha, H. Peltola, T. Ruff, et al. Vaccination greatly reduces disease, disability, death and inequity worldwide. *Bulletin of the World Health Organization*, 86(2):140–146, 2008.
2. L. Au, A. Oster, G. Yeh, J. Magno, H. Paek, et al. Utilizing an electronic health record system to improve vaccination coverage in children. *Appl Clin Inform*, 1(3):221–231, 2010.
3. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.
4. S. Barlas. Fda weighs updating its bar-code mandate: Hospital pharmacies worry about implementation. *Pharmacy and Therapeutics*, 37(3):162, 2012.
5. A. Katib, D. Rao, P. Rao, K. Williams, and J. Grant. A prototype of a novel cell phone application for tracking the vaccination coverage of children in rural communities. *Computer Methods and Programs in Biomedicine*, 2015.
6. F. Michel, J. Montagnat, and C. Faron-Zucker. A survey of rdb to rdf translation approaches and tools. 2014.
7. Monika Solanki and Christopher Brewster. Representing Supply Chain Events on the Web of Data. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE) at ISWC*. CEUR-WS.org proceedings, 2013.
8. M. Solanki and C. Brewster. EPCIS event based traceability in pharmaceutical supply chains via automated generation of linked pedigrees. In Peter Mika et al., editor, *Proceedings of the 13th International Semantic Web Conference (ISWC)*. Springer-Verlag, 2014.
9. M. Solanki and C. Brewster. Monitoring EPCIS Exceptions in linked traceability streams across supply chain business processes. In *Proceedings of the 10th International Conference on Semantic Systems(SEMANTiCS)*. ACM-ICPS, 2014.
10. D. Thornton, H. Mwanyika, D. Meek, and U. Kreysa. Tanzania leading the way with barcodes on vaccine packaging. *OPTIMIZE*, 2013.
11. M. Zaffran, J. Vandelaer, D. Kristensen, B. Melgaard, P. Yadav, K. Antwi-Agyei, and H. Lasher. The imperative for stronger vaccine supply and logistics systems. *Vaccine*, 31:B73–B80, 2013.