J Digit Imaging (2016) 29:460–465
DOI 10.1007/s10278-016-9867-z

Using Object Storage Technology vs Vendor Neutral Archives for an Image Data Repository Infrastructure

Brian Bialecki[1] · **James Park**[2] · **Mike Tilkin**[3]

Published online: 12 February 2016

Abstract The intent of this project was to use object storage and its database, which has the ability to add custom extensible metadata to an imaging object being stored within the system, to harness the power of its search capabilities, and to close the technology gap that healthcare faces. This creates a non-disruptive tool that can be used natively by both legacy systems and the healthcare systems of today which leverage more advanced storage technologies. The base infrastructure can be populated alongside current workflows without any interruption to the delivery of services. In certain use cases, this technology can be seen as a true alternative to the VNA (Vendor Neutral Archive) systems implemented by healthcare today. The scalability, security, and ability to process complex objects makes this more than just storage for image data and a commodity to be consumed by PACS (Picture Archiving and Communication System) and workstations. Object storage is a smart technology that can be leveraged to create vendor independence, standards compliance, and a data repository that can be mined for truly relevant content by adding additional context to search capabilities. This functionality can lead to efficiencies in workflow and a wealth of minable data to improve outcomes into the future.

Keywords Image database · Clinical application · Information storage and retrieval · PACS integration · Image feature enhancement · Information visualization · Data extraction · Infrastructure · Integrating Healthcare Enterprise (IHE) · Imaging informatics · Workflow re-engineering · Enterprise PACS · Information management · PACS planning

✉ Brian Bialecki
bbialecki@acr.org

[1] CIIP, American College of Radiology, 1818 Market Street Suite 1720, Philadelphia, PA 19103, USA

[2] Hitachi Data Systems, Santa Clara, CA, USA

[3] American College of Radiology, 1891 Preston White Drive, Reston, VA 20191, USA

Background

Finding innovative ways to make image storage and access more efficient, less costly, and more extensible are imperatives facing all organizations involved in complex imaging tasks. During an evaluation of current storage solutions the following question was encountered: "Can commercial object storage, with custom metadata tagging, compete in a healthcare environment as a VNA (Vendor Neutral Archive) alternative in certain use cases?" Rich tool sets developed by object storage vendors, such as XML-based (Extensible Markup Language) customizable annotation or extensible metadata [1], offer the ability to use these common storage platforms for advanced image and correlative data management. Embedding this information with a standard-based focus creates a single repository for images as well as their artifacts without changing the current technology or workflow. Object storage will allow for unlimited scalability in terms of the amount of data, individual data object size, and object complexity. Adoption of integrated image management solutions in healthcare using more general storage toolsets will promote data sharing and mining in line with other industries. For example, currently, it is easier to get money from personal bank accounts, securely and from anywhere in the world, than it is for a physician, who is providing lifesaving treatment, to access relevant information when making decisions for patient care and well-being. Through the use of standard-based extensible metadata, cloud-friendly technology and support for APIs (application programming interfaces), object storage promotes security, flexibility, and ease of

integration, making pertinent information available when and where it is needed most. It is noted that object storage cannot easily support workflow on its own merit, but the non-disruptive technology and the features that it introduces to the industry will enable healthcare to close the technology gap when compared to other industries. These gaps are evident in the duplicative exams resulting from the struggle to share data, as well as the daily use of legacy technologies for critical tasks.

Traditionally, VNAs are archives built on standards and are capable of interfacing with various vendors for acquisition, viewing, and workflow. They can be used to store both DICOM (Digital Imaging and Communications in Medicine) and non-DICOM content for multiple departments within an enterprise or federated architecture. A VNA also allows for data transformation and is capable of master patient indexing for reconciliation and queries across institutions, systems, and data types. "Characteristic for a VNA is that it provides a patient-centric approach that transcends upgrades and changes of the different viewing, acquisition, and workflow management components as they should be interchangeable without having to migrate, convert, or change the data formats or interface of the VNA" [2]. Consumer systems that can process data stored in any standard-based repository have been referred to as archive neutral vendors [3].

Object storage separates file metadata from the data it supports. The files themselves become objects referenced by metadata and the metadata is stored in a database. This concept sparked the comparison to PACS (Picture Archiving and Communication System) and became the impetus for this project. Throughout this paper, the metadata database will be referenced as the place where annotations and extensible functionality reside for the simplicity of documenting the concept. However, in essence this is not truly a traditional database such as Oracle or SQL, but the built-in metadata functionality of object storage.

As object store technology advances, and where traditionally healthcare lags behind, adopting a commercial business solution for big data storage becomes a more viable and tool-rich option. Object stores increasingly support functionality such as CIFS (Common Internet File System) and NFS (Network File System), which are staples of our current PACS environments. This allows object storage technology to be introduced into current workflows and environments without disruption, where it will be seen by current PACS archives, workstations, or VNAs as supported storage. These technologies also natively support RESTful (Representational State Transfer) and other Internet-based protocols, allowing for incorporation with more enterprise level applications and easier integration with developing technologies. Object storage platforms allow for the writing of data using one protocol and reading of the same data with another. Other potential advantages of such technology include object deduplication and scalability. Object stores hash the block contents and compare those hashes when storing new content. In this way, as copies of images move around the enterprise deduplication can be achieved as a function of the object store itself. When installed in a manner where object storage is managing all enterprise image caches and repositories, it becomes highly effective as only a single copy of the image is stored in the system while presenting each cache and logical drive what would appear to be a local copy, represented as a pointer to the real image file. Additionally, as data gets larger, pathology and genomics files for example, and we begin to encounter the flood of big data, the current CIFS/NFS [4] technologies limit flexibility and scalability. Object-based storage addresses these limitations.

Object storage vendors have begun to allow custom extensible metadata annotations to be added to the object metadata being placed in their databases through the Object Store Alliance [1]. When accessed through API level searches, the addition of these custom metadata elements begins to approach the needed functionality of enterprise imaging applications and VNA level storage, while providing HITRUST (Health Information Trust Alliance) [5] compliant security. The project concept was to have multiple custom metadata elements linked to a DICOM and/or non-DICOM object. For DICOM objects, this would be an XML representation of the DICOM header as well as other correlative data artifacts. Once objects have custom metadata placed around them, advanced searches can be done against this metadata content in order to locate a specific stored object. RESTful interfaces can then be used to present the objects for rendering with thin or zero footprint viewers through APIs and the IHE (Integrating the Healthcare Enterprise) Invoke Image Display [6] profile. These viewer technologies, such as Claron's Nil [7] and Client Outlook's eUnity [8], have increasingly seen approval for diagnostic purposes by the Food and Drug Administration (FDA).

The use case of object storage providing enhanced VNA functionality without the DICOM communications overhead was tested, which leveraged the advantages of the object storage technology platform for providing security, being infinitely scalable and having the flexibility of extensible metadata annotations. This testing was accomplished while providing non-disruptive interfaces to legacy healthcare applications and workflows. This was in no way an attempt to take away from any standard. In fact, strict adherence to the object formatting based on the DICOM standard and others will be key to healthcare's continued success and the success of this project. The addition of this technology to an enterprise's current infrastructure, along with the powerful features that are made available, will enhance healthcare's ability to close the technology gap it continually faces.

Case Presentation

It is well documented that a VNA will be able to store, find, and retrieve image data based on the metadata provided within

each DICOM image file. DICOMweb [9] has opened up RS (RESTful services) to the DICOM community and these too are supported by the VNA vendors. STOW (Store over the web), QIDO (Query based on ID for DICOM Objects) and WADO (Web Access of DICOM Objects) allow for store, find, and retrieve. Alternatively, object storage allows for the Internet-based file transfer protocols of PUT and GET to function as the equivalent of store and retrieve. These PUT and GET actions were performed as substitutes for DICOMweb commands, with the data encrypted while in transit. The creation of these standard services allowed for the success of this project as it enabled a standard abstraction of the DICOM metadata from the image pixel data, as well as the proliferation of downstream systems that can process standard-based objects on their own. Metadata annotations that were added for each imaging file provided a database for searches using Solr [10] formatting, and the object store API used to pass them was equivalent to a query. This is where the power of object storage was realized. By extending the metadata customization beyond the DICOM image metadata to include additional ancillary artifacts, searches could be more contextual. As a result, more complex queries can be made that cross departments, yet have relevance to the stored imaging object and allow for powerful data mining.

The first step in testing these APIs and the use case was to determine the types of annotations and ancillary content that would be applied as extensible metadata to the stored image objects. The goal was to use standards that already existed so that this implementation was practical in today's healthcare environment. Custom annotation fields were written to via API within the object storage database. Up to ten (10), two (2) GB XML annotations could be written, updated, and searched per storage object. This is currently a limit of the object storage code as there have been no use cases found where more than five (5) have been used. Although there is no theoretical limit to this number, there is willingness from the vendor to extend this to cover use cases as they present themselves. HL7 V3 CDA (Clinical Document Architecture) [11] standards allow for History and Physicals, Operative Reports, Clinical Visit Summaries, and Lab Reports to be represented within the object annotations. These items can be linked as ancillary artifacts of imaging as the reason for order or the result of an image guided biopsy. DICOM SR (Structured Report) has an XML representation [12] defined by the standards, which is a direct result of the radiologists' visualization of the image. In the WADO-RS documentation [13], the retrieval and construction of DICOM metadata is defined as an XML structure. There are also emerging standards such as the caBIG AIM (Annotation Image Markup) project [14] and vendor workstations that have reproducible reporting schema where findings are represented in XML. As these custom annotations were added to the objects, they were defined by templates called content classes by the object storage which made the searches of these annotations more definable and faster than looking at text strings. This allowed for targeted searches of data within a defined XML structure as opposed to processing the entire document as text, which then requires parsing the complete contents for each query. All the annotations were added to the object store database, while the applications using the storage continued to see it as raw storage presented as CIFS or NFS and without any disruption to the end users. These files were accessible normally through applications graphical user interfaces (GUI), as well as directly from the storage, in context with the additional searchable and standard-based data that was populated in the metadata annotation fields.

Unique images were then gathered for testing and the metadata was extracted from the parsed DICOM file using ImageJ [15] to create the XML representation of the DICOM header metadata. In addition, relative CDA documents were added to each of the image files. These CDA documents were created from the clinical encounter that took place in which the ordering physician placed the imaging order, allowing for data mining regarding the clinical presentation of the patient and the reason for the exam. Subsequently, the CDA documents were validated using the NIST tools [16] to ensure they were compliant with the templates that had been setup as content classes. AIM XLMs [14] that were created during a review of these images using Daniel Rubin's ePAD application [17] were also used. Finally, annotations from the primary reviewing workstation, which provided a report in XML format, were included. This report had a schema that was validated and was provided by the vendor as being standard within their environment. The image files were placed into the object store using curl scripts [18] or directly by applications, while the annotations were all placed into the database using curl scripts. Once the system had been populated with both image files and correlative annotations, the consumption and search capability of the system was tested. Figure 1 shows the data sources and test system structure used for this project.

Normally, consuming the data that was stored within the single test environment would require a user to execute queries for information from each of the many disparate systems in their clinical environment, then manually correlating this information. A user wishing to return all CTs done within a specific date range at their institution could go to their PACS and search DICOM tags (0008,0060) Modality and (0008, 0020) Study Date. The approach taken to test and compare searches was that of querying the source system and then the test system. The PACS returns were compared to direct API queries of the object storage for the same values from the annotation XMLs, specifically the XML representation of the DICOM header. The data returned matched in both instances. A user would then be able to go to another system, such as their EMR (Electronic Medical Record), retrieve patients who presented with appendicitis during this same date

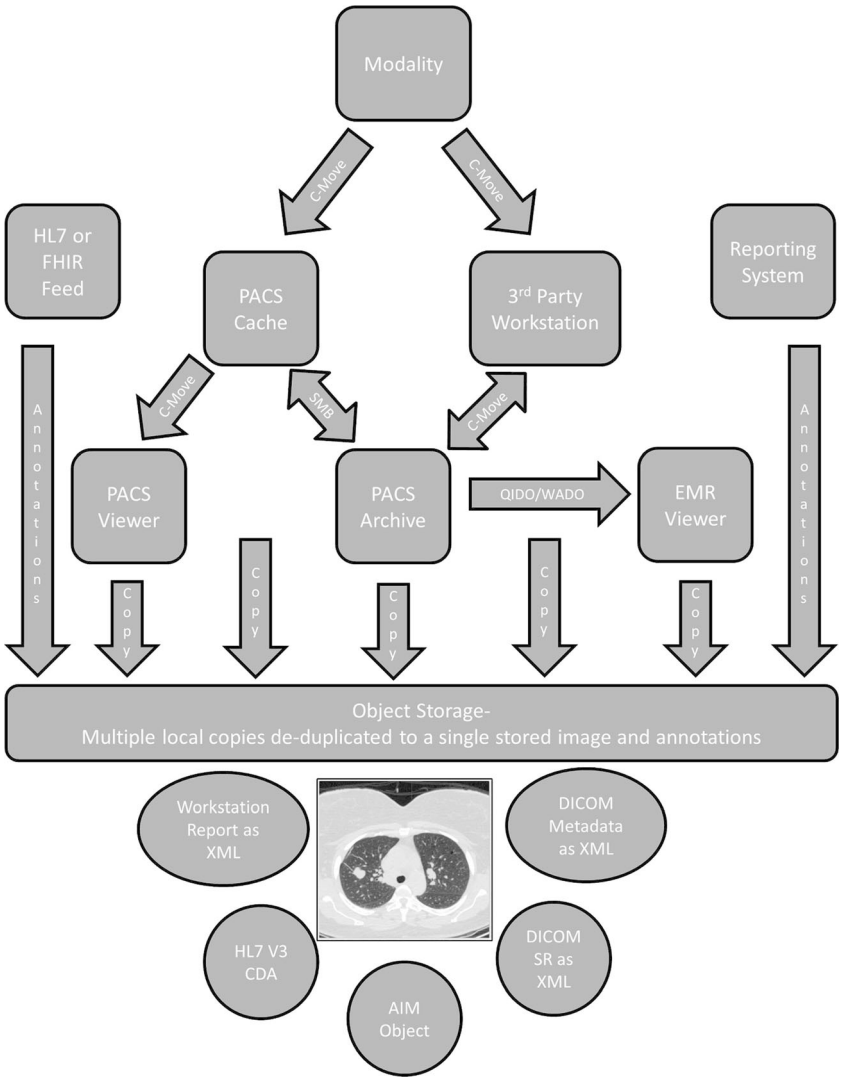

Fig. 1 System structure with inputs

range, and compare the results to the patients with CTs to determine which patients had a CT based on their clinical presentation. This query was completed using the search APIs to the object store [19] annotations, and all patients who had imaging performed were returned. The resulting list from the API search of the test system contained fewer patients than the EMR search as the presence of multiple data sources within each image's searchable annotations enabled tighter constraints to be used. When the EMR results were cross-referenced with the PACS results, the subsets of these returns again matched. Figure 2 shows the data structure and the process for consuming the contents of the test system. Queries were also completed in the reviewing system as well as the reporting system, and this information was compared to the API queries of the object storage. Each time, the overlap of information was identical. This in itself simply proved that when populated properly, the object storage was capable of returning equivalent data to the comparison systems. While this is true, the data had now been federated and stored with the correlating images, allowing for a much more complex search and targeted return. Using the context classes that were built in the object storage, a single search of the annotations that encompassed all of the parameters of each of the four systems' individual searches was completed and returned a specific result. Figure 3 is an example of a federated query generated and executed against the test system using curl.

The following query was asked of the object storage: Return any CT that was completed between 20120401 and 20141231 where the patient was clinically presented with appendicitis, but only include images where a liver lesion was marked up during the review and there were additional markups made that were defined as lymph nodes. To obtain this information, the following locations were searched: DICOM header in XML for the CT and study date range, clinical notes CDAs for appendicitis, review workstation's XML report for a lesion defined in the body part "Liver," and the AIM markups for the RADLEX [20] code for lymph node. This

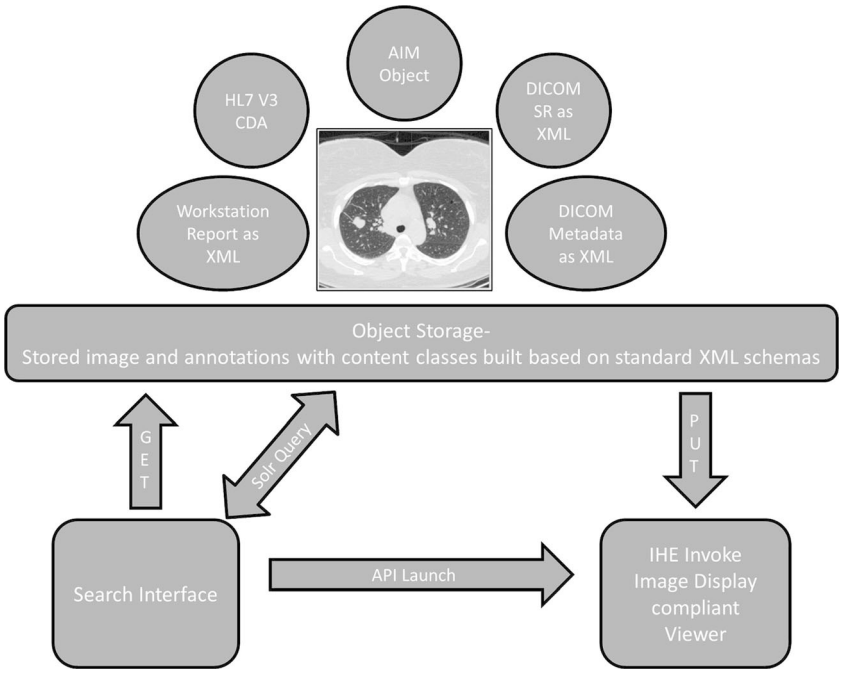

Fig. 2 Data structure for query and consuming

resulted in the path to a unique image file that met these parameters being returned through the API, which was then used in a batch file where a GET command brought that specific image locally over the Internet in a secure transaction. The image was subsequently launched using the IHE Invoke Image Display [6] profile and displayed for the user to visualize and interact with.

```
<queryRequest>
        <object>
        <query>+(namespace:"ns.demo")+(CodingSchemeDesignator:RadLex)
                +(CodeValue:RID13296)</query>
        <verbose>true</verbose>
        </object>
        <object>
        <query>+appendicitis</query>
        <verbose>true</verbose>
        </object>
        <object>
        <query>+(namespace:"ns.demo")+(elmTag:00080060)+(ValData:CT)
                </query>
        <verbose>true</verbose>
        </object>
        <object>
        <query>+(namespace:"ns.demo")+(elmTag:00080020)
                +(StudyDate:[2012-04-01T00:00:00Z TO *])
                +(StudyDate:[* TO 2014-12-31T23:59:59Z])</query>
        <verbose>true</verbose>
        </object>
        <object>
        <query>+(namespace:"ns.demo")+(Lesion_Organ:liver)</query>
        <verbose>true</verbose>
        </object>
</queryRequest>

echo "Search All Annotations as Federated"
curl -k -H "Authorization: HCP dXN1cjE=:bed128365216c019988915ed3add75fb" -H
    -H "Content-Type: application/xml" -H "Accept: application/xml" -d
    @Federated.xml "https://demo.hcp.hcpdomain.lab/query?prettyprint"
```

Fig. 3 Federated query

Discussion

A search engine this powerful, with federated data that can be provided in such a context, sparks ideas for many further use cases. Since this technology can be added to existing infrastructure in a non-disruptive fashion, it will allow for easy adoption and value-added propositions. The ability to find truly relevant priors is a potential use case to tackle. Algorithms [21] for finding priors can be developed on a much more complex set of criteria. Presenting data that is much more correlated based on the patient's presentation and previous findings, instead of body part alone, can be used to enhance workflow for the reader. This can also ensure that with the overwhelming amount of correlative data available, the reader can be made aware of truly relevant priors that might otherwise be overlooked in the volume of data being generated, in addition to the gains in efficiency. The long-term value can only be imagined at this time. To be able to ask questions once a history has been developed, and then return to the past to test ideas against live data, empowers researchers with a wealth of new data sources.

Future work on this project will be needed to create an algorithm for assigning and selecting data for annotation to images, although it is a consideration that this could be site specific based on use case. It would also be useful to deploy Natural Language Processing (NLP) [22] to create the Solr searches in a much more user-friendly interface. Finally, testing implementations with VNA vendors to add these annotations to the objects they store would create a much faster population strategy that could be more easily deployed in a broader user base.

Although this project initially began as a test on the feasibility of object storage vs. VNA, and in some specific use cases it does appear that it is truly a viable alternative, a more complementary approach may be in line with other use cases. It appears that VNA technology may be an ideal complement for the implementation and population of this technology in a fairly short period, while enabling users to take advantage of direct access to their data in an unhindered fashion. This would allow for the maintenance of legacy workflows while empowering users with the latest in technology, giving them greater flexibility and control over the data, and beginning to level the technology gap that healthcare finds itself in.

A project that will leverage this technology is now underway as part of a larger data warehousing and archive federation effort. Additional tools are currently being developed to incorporate and expand the interfaces, APIs, and overall capabilities on both the vendor and consumer sides of this technology stack.

References

1. "Object Storage Alliance." Neuralytix. Available at http://www.neuralytix.com/osa. Accessed 18 May 2015.
2. Oosterwijk, Herman. "What Is a VNA, Anyway?" What Is a VNA, Anyway? By Herman Oosterwijk, President, OTech Inc. Available at http://www.himss.org/files/HIMSSorg/content/files. Accessed 18 May 2015.
3. Yeager, David. "Vendor-Neutral Archive or Archive-Neutral Vendor?" Vendor-Neutral Archive or Archive-Neutral Vendor? Radiology Today, n.d. Web. 2015.
4. "Difference Between NFS and CIFS | Difference Between | NFS vs CIFS." Difference Between NFS and CIFS | Difference Between | NFS vs CIFS. N.p., n.d. Web. 2015.
5. HITRUST. Health Information Trust Alliance. Available at https://hitrustalliance.net. Accessed 28 Dec 2015.
6. Committee, IHE Radiology Technical. IHE_RAD_Suppl_IID_Rev1.2_TI_2015-04-21 (n.d.): n. pag. Ihe.net. IHE. Web. 2015.
7. "Claron Receives FDA 510k Clearance For NilRead Diagnostic Zero-Footprint Medical Image Viewer." Health IT Outcomes. Available at http://www.healthitoutcomes.com/doc/claron-nilread-diagnostic-zero-footprint-medical-image-viewer-0001 Accessed 28 Dec 2015.
8. "Client Outlook awarded FDA 510(k) Class II Medical Device Clearance for eUnity" Client Outlook Inc. Available at http://www.clientoutlook.com/press_release/client-outlook-awarded-fda-510k-class-ii-medical-device-clearance-for-eunity. Accessed 28 Dec 2015.
9. "DICOMweb." DICOMweb. Available at http://dicomweb.hcintegrations.ca/#/home. Accessed 18 May 2015.
10. "Solr Features." Apache Solr. Apache. Available at http://lucene.apache.org/solr/features.html. Accessed 18 May 2015.
11. HL7 International. "CDA Release 2." Available at http://www.hl7.org/implement/standards. Accessed 18 May 2015.
12. "Dsr2xml: Convert DICOM SR File and Data Set to XML." OFFIS DCMTK. Available at http://support.dcmtk.org/docs/dsr2xml.html. Accessed 18 May 2015.
13. "6.5 WADO-RS Request/Response." 6.5 WADO-RS Request/Response. Available at http://medical.nema.org/dicom/2013/output/chtml/part18/sect_6.5.html. Accessed 18 May 2015.
14. Channin, David S. et al. "The caBIGTM Annotation and Image Markup Project." Journal of Digital Imaging: the official journal of the Society for Computer Applications in Radiology 23.2 (2010): 217–225. PMC. Web. 2015.
15. "ImageJ." ImageJ. Available at http://rsb.info.nih.gov/ij/index.html. Accessed 18 May 2015.
16. "CDA Guideline Validation." CDA Guideline Validation. Available at http://cda-validation.nist.gov. Accessed 18 May 2015.
17. "ePAD: A Cross-Platform Semantic Image Annotation Tool" RSNA. Available at http://www.researchgate.net/publication/266104399_ePAD_A_Cross-Platform_Semantic_Image_Annotation_Tool. Accessed 18 May 2015
18. "Documentation Overview." CURL. Available at http://curl.haxx.se/docs/. Accessed 18 May 2015.
19. Systems, Hitachi Data. "HCP Using HS3 API." (n.d.): n. pag. Hitachi Data Systems. Web. 2015.
20. "RadLex Term Browser." RadLex Term Browser. Available at http://www.radlex.org. Accessed 18 May 2015.
21. "Understanding the Principles of Algorithm Design - Tuts+ Code Tutorial." Code Tuts+. N.p., n.d. Web. 2015.
22. "OpenNLP." Solr Wiki. Available at https://wiki.apache.org/solr/OpenNLP. Accessed 18 May 2015.