Open Source Clinical NLP – More than Any Single System

James Masanz, MS[1], Serguei V. Pakhomov, PhD[2], Hua Xu, PhD[3], Stephen T. Wu, PhD[1], Christopher G. Chute, MD DrPH[1], Hongfang Liu, PhD[1]
[1]Mayo Clinic, Rochester, MN, [2]College of Pharmacy and Institute for Health Informatics, University of Minnesota, [3]School of Biomedical Informatics in The University of Texas Health Science Center at Houston

Abstract

The number of Natural Language Processing (NLP) tools and systems for processing clinical free-text has grown as interest and processing capability have surged. Unfortunately any two systems typically cannot simply interoperate, even when both are built upon a framework designed to facilitate the creation of pluggable components. We present two ongoing activities promoting open source clinical NLP. The Open Health Natural Language Processing (OHNLP) Consortium was originally founded to foster a collaborative community around clinical NLP, releasing UIMA-based open source software. OHNLP's mission currently includes maintaining a catalog of clinical NLP software and providing interfaces to simplify the interaction of NLP systems. Meanwhile, Apache cTAKES aims to integrate best-of-breed annotators, providing a world-class NLP system for accessing clinical information within free-text. These two activities are complementary. OHNLP promotes open source clinical NLP activities in the research community and Apache cTAKES bridges research to the health information technology (HIT) practice.

1. Introduction

Rapid growth in the clinical implementation of large electronic medical records (EMRs) has led to an unprecedented expansion in the availability of dense longitudinal datasets for clinical and translational research[1, 2]. This growth is being fueled by recent federal legislation that provides generous financial incentives to institutions demonstrating aggressive application and "meaningful use" of comprehensive EMRs[3-5]. Efforts are already underway to link these EMRs across institutions and to standardize the definition of phenotypes for large scale studies of disease onset and treatment outcome, specifically within the context of routine clinical care[6-8]. A well-known challenge in the secondary use of EMR data for clinical and translational research is that much of detailed patient information is embedded in narrative text. It is a very time-consuming and costly process to manually extract information from clinical records[9-11]. Researchers have used NLP systems to identify clinical syndromes and common biomedical concepts from radiology reports, discharge summaries, problem lists, nursing documentation, and medical education documents[12-18].

Different NLP systems have been developed at different institutions and used to convert clinical narrative text into structured data that may then be used for other clinical applications and studies. The systems include MedLEE[19, 20], MetaMap[21], KnowledgeMap[22], Apache cTAKES[23,24], and HiTEX[25]. Successful stories in applying NLP to clinical and translational research have been reported widely, ranging from identifying patient safety occurrences[26] to facilitating genomics research such as gene-disease association analysis and pharmacogenomic studies[15-18, 27, 28]. However, the lack of interoperability of NLP systems limits the full potential of applying NLP for clinical and translational research.

Multiple national-level informatics initiatives aim to enable the secondary use of EMRs for clinical and translational research but lack dedicated effort in addressing interoperability of existing NLP systems. The i2b2 (Informatics for Integrating Biology and the Bedside) center, one of the NIH roadmap centers, has developed a scalable informatics framework for clinical and translation research. One NLP system, HiTEX, has been distributed through the i2b2 platform as an optional component, and Apache cTAKES has become the platform of choice for i2b2. The NLP activities in iDASH (integrating Data for Analysis, Anonymization, and Sharing), another roadmap national center funded by NIH, have been focusing on developing an NLP ecosystem to i) develop and disseminate shareable NLP tools and annotated data, ii) determine what gaps exist between current efforts and an ideal software/data ecosystem, and iii) outline needs for integrating our efforts, data, and tools. Another informatics initiative, SHARP Area 4 project headed by Mayo Clinic has focused on building infrastructure tools including NLP through Apache cTAKES for high throughput phenotyping. The Consortium for Healthcare Informatics Research (CHIR) has been adopting the best-of-breed NLP methods to unlock clinical information available as free text in the VA EMR system. All those efforts promote uses of NLP for clinical investigation, point-of-care applications, and support for decision support[29,30].

We believe there are two factors causing the lack of interoperability among existing NLP systems: i) closed source development and ii) lack of standardization of NLP data models and exchange format. The recent development and adoption of open source engineering framework architectures such as GATE (General Architecture for Text Engineering)[31] and UIMA (Unstructured Information Management Architecture)[32] in the research community has enabled scalable and modular development of tools for processing unstructured information (text included). In the biomedical domain, several groups have successfully implemented NLP platforms based on UIMA or developing wrappers around existing open source NLP tools[33-35].

In this paper, we present two ongoing activities to promote open source interoperable clinical NLP research and development. One is the development of the infrastructure for The Open Health Natural Language Processing (OHNLP) Consortium and the other is the release of cTAKES as a top-level project within the Apache Software Foundation. In the following, we first provide some background information. We then present the ongoing activities in OHNLP and cTAKES in more detail.

2. Background and Motivation

The Open Health Natural Language Processing (OHNLP) Consortium was founded in 2009 to release open source clinical NLP tools developed under UIMA. Two UIMA-based NLP systems with distinct NLP data models were initially contributed to OHNLP. One is cTAKES (clinical Text Analysis and Knowledge Extraction System) and the other is MedKAT/p[36]. The two systems were not able to directly communicate with each other; even though both were built upon the UIMA framework. The reasons for this were:

- **Independent development with different purposes** - cTAKES and MedKAT/p were developed at different times and with different aims, even though there was some overlap in the people involved in both. cTAKES was initially developed to process clinical notes, while MedKAT/p was written to process pathology reports. This resulted in the data models have different focuses, and not being aligned.
- **Java naming conventions** - The two systems were written in Java but not owned by a single institution. Java naming conventions generally use institution names within the source code (using com.ibm and edu.mayo within the Java package names); the names used within the two data models (i.e., UIMA type systems) had no chance of being the same. For the two systems to share data seamlessly requires either changing the source code of one or both of them, or writing some interfacing code.

When all components for a UIMA pipeline come from a single development team, the team can agree on the type system, and data can flow seamlessly between components. When components are developed by different teams, often some interfacing code must be written to convert data from one structure or format to another. Apart from the data type disparities, interoperability between NLP systems may be affected by use of different linguistic theories and approaches to representing linguistic entities. For example, one NLP system may rely on the original Penn Treebank[37] tags for part-of-speech tagging, whereas another system may use a modification of Penn Treebank or an entirely different tag set altogether. This is a particularly problematic interoperability issue because the disparity between tag sets may not necessarily lead to an overt and easily detectable system failure, but instead result in unpredictable and possibly less accurate system performance. Therefore, raising the awareness in the NLP community of these issues is an important first step toward creating interoperable NLP systems and components. The interoperability must exist between NLP system designers and developers first in order for the interoperability between NLP systems to follow.

There are multiple activities underway working toward interoperable clinical NLP. Specifically, under the SHARPn project, an NLP common data model was developed to be more comprehensive while still allowing for extensions[38]. That data model has been adopted by Apache cTAKES as well as multiple open source NLP tools released under OHNLP including MedTagger, MedXN, and MedTime.

To create community awareness of existing informatics tools, the ORBIT site (hosted by National Library of Medicine) catalogs informatics software, knowledge bases, data sets and design resources[39]. However, the ORBIT site does not provide any hosting services for projects themselves. Developers of informatics software who wish to release the software open source must decide upon a code repository, a site for documentation, how to provide a place for questions to be asked, etc. The iDASH Natural Language Processing (NLP) Ecosystem provides a link to the ORBIT registry as well as a virtual machine image that you can download which contains several NLP tools[40]. It intends to also host source code; however it does not yet host source code, nor does it provide a way to download the tools individually.

The Apache Software Foundation (ASF) - The ASF provides Wiki space, mailing lists, options for issue tracking (JIRA and Bugzilla), worldwide mirroring of distribution website, press releases for new projects, and expertise in a variety of areas such as licensing and trademarks. The ASF's mission is to provide software for the public good. The ASF however requires its projects have a diverse set of committers. The ASF's guide to creating a proposal for a new Apache project states "Apache is interested only in communities. Candidates should start with a community and have the potential to grow and renew this community by attracting new users and developers."(41) Tools are typically not shared through ASF until a certain level of maturity has been reached.

3. Towards Community-Driven Open Source Clinical NLP

Figure 1 illustrates the history as well as our thought towards open source clinical NLP which was pioneered by the establishment of OHNLP. Initially OHNLP was home to two clinical NLP systems – cTAKES and MedKAT/p. Other systems have since been contributed to OHNLP, and cTAKES has moved on to the ASF and become Apache cTAKES. OHNLP has added to its original mission – it will be not only a source of information for clinical NLP software outside of OHNLP, but it will also be the source of interfaces that address interoperability between existing NLP software. Therefore, OHNLP can be viewed not only as a home for mature medical NLP tools but also as an innovation incubator for projects in early experimental stages of development

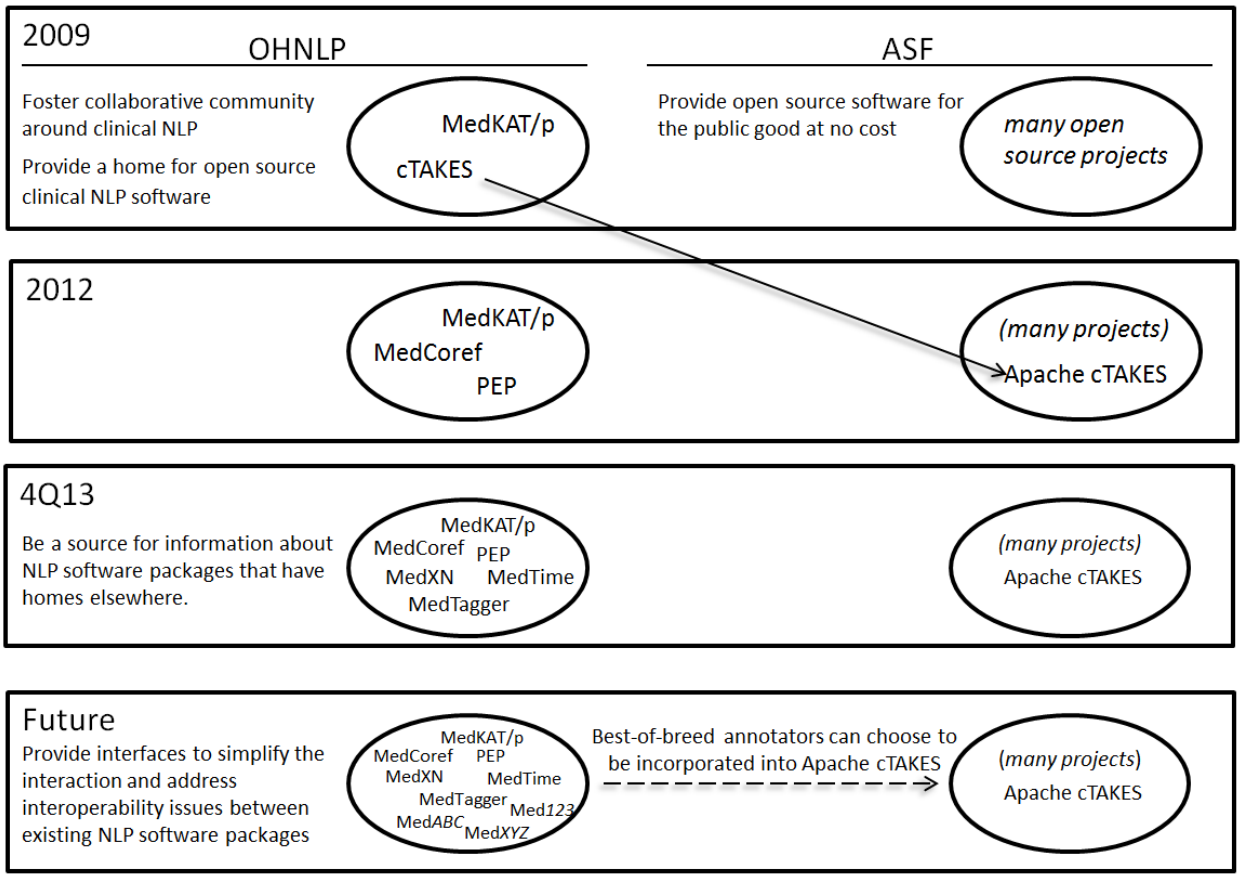

Figure 1. OHNLP and Apache cTAKES Timeline

3.1. Apache cTAKES

After an initial innovation incubation phase at OHNLP, cTAKES has migrated to Apache cTAKES, which strives to be a world-class clinical Natural Language Processing (NLP) system, containing best-of-breed annotators. It is open source and the Apache cTAKES community welcomes contributions. Apache cTAKES is modular and expandable at the information model and method level. As an Apache project, it is licensed under the Apache License, Version 2.0 by the Apache Software Foundation (ASF), which is how the ASF licenses all its current projects .

Apache cTAKES aims to be modular and expandable at the information model and method level and can be used in a great variety of retrievals and use cases. The Apache cTAKES community is committed to best practices and R&D (research and development) by using cutting edge technologies and novel research and facilitating the translation of the best performing methods into the project.

Apache cTAKES has incorporated and intends to continue to incorporate best-of-breed annotators. The community is multi-institutional or non-institutional – not dependent upon any single institution for funding. One goal is to align with industry/community standards and conventions.

Coinciding with cTAKES becoming an Apache project, work was done to generalize the type system (i.e., data model). However there are still many different type systems in use by different tools and systems. Work to have a basic common type system across several systems is under discussion – it is only in the planning stage and only addresses a few systems.

3.2. The OHNLP Consortium

The mission of the current OHNLP Consortium consists of three main objectives:

1. Provide a home for existing and emerging clinical NLP software.

2. Provide a gateway to NLP software packages that reside in other repositories.

3. Provide interfaces to simplify the interaction of various NLP packages and address interoperability issues between existing NLP software packages; allowing seamless information exchange among them.

Be a home for open source clinical NLP software. Existing open source repositories such as the Sourceforge provide open source software developers a place to host source code, publish releases, maintain documentation, and track issues. Creating a new Sourceforge project does not provide the visibility that comes with being associated with the OHNLP Consortium. For those tools/systems looking for a home, OHNLP can be the place for the tool/system to live, and for a community to develop around it. OHNLP has a dedicated website with a Wiki, and a project (https://sourceforge.net/projects/ohnlp/) at Sourceforge with:

- a repository for the code (an SVN instance)
- issue (bug) trackers
- forums
- release staging and publishing

While the OHNLP Consortium encourages the use of the Apache License, OHNLP does not comprise a single system, therefore the consortium does not prohibit individual systems or tools that are contributed to it from being licensed under the GNU General Public License (GPL) or other licenses that would be problematic for Apache. OHNLP can provide a home for these tools/systems. OHNLP can be the long term home of an NLP tool/system, or it can be a starting place of a tool/system that moves on to somewhere else such as the ASF. There is no need for a tool's owner to decide ahead of time.

Be a registry for existing clinical NLP systems, tools, and resources. OHNLP is intended as a hub for clinical NLP resources and tools; aiming to facilitate sharing of resources and of software. And beyond simply sharing software packages (binaries), one of OHNLP's goals is to encourage the developers of clinical NLP software to release their code as open source. If an institution requires others to sign a data use agreement (DUA) before accessing a shared resource, OHNLP can be the home for the DUA request forms and contact information.

Promote interoperability by a common data model and address interoperability among existing NLP systems. There is a need for existing NLP tools/systems to be able to share data (interoperate) with each other, including Apache cTAKES. But there is much work to be done to make existing clinical NLP tools interoperate better. Once tools and systems are established it is often difficult to find the time to work on interoperability. A common data model (OHNLP types) built upon the SHARPn common data model is currently available at OHNLP. Its intent is not to dictate the use of specific data types and structures but to provide a starting point and make it easier for developers to consider interoperability with other systems as one of their objectives. Thus, developers can extend the common data model for their specific systems or build their systems in a way that will not preclude future integration with other tools in the OHNLP initiative. There are two options for improving how existing systems interoperate – develop a common type system/data structure and rework the existing systems to use them, or develop

interfaces to allow the existing systems to interoperate. Developers should be able to exercise either of the options for improving the interoperability. For example, the NLP team at Mayo Clinic recently reworked their NLP systems including MedTagger, MedTime, and MedXN to adopt the common data model. Currently, OHNLP is in the process of developing interfaces to multiple open source clinical NLP systems including Apache cTAKES, MedEx, BioMedICUS, MedLEE, MetaMap, and HiTEX.

4. Conclusion

We have described two ongoing activities for open source clinical NLP. The OHNLP Consortium provides to the clinical NLP community resources that are not found elsewhere under a single umbrella. OHNLP provides visibility to NLP software and OHNLP infrastructure includes a Wiki, a registry of related software, and a place where a tool or system without an established diverse development community – including tools developed by a single author – can be hosted. Meanwhile, Apache cTAKES aims to provide best-of-breed NLP modules to the community and facilitates the translation of research into practice.

Future work would include the continuous development of both OHNLP and Apache cTAKES and the collaboration with other initiatives to advance open source clinical NLP.

5. Acknowledgements

We would like to thank and acknowledge the work of Anni Coden, Ph.D. Michael Tanenblatt, Igor Sominsky, et al. at IBM as well as Guergana Savova, Ph.D. and others involved in founding the OHNLP Consortium. We would also like to thank and acknowledge the work of Guergana Savova and Pei Chen of Boston Children's Hospital and Harvard Medical School as well as others in transitioning cTAKES to Apache cTAKES. The work on OHNLP is funded in part by R01 GM102282. The work on Apache cTAKES was funded in part by the SHARPn (Strategic Health IT Advanced Research Projects) Area 4: Secondary Use of EHR Data Cooperative Agreement from the HHS Office of the National Coordinator, Washington, DC. DHHS 90TR000201.

References

1. Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. New England Journal of Medicine. 2010;362(3):192-5.
2. Jha AK, DesRoches CM, Campbell EG, et al. Use of electronic health records in US hospitals. New England Journal of Medicine. 2009;360(16):1628-38.
3. Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. N Engl J Med. Jan 21;362(3):192-5.
4. Secretary Sebelius Announces Final Rules To Support 'Meaningful Use' of Electronic Health Records. US Department of Health and Human Service 2010 [cited 2010 10/15]; Available from: http://www.hhs.gov/news/press/2010pres/07/20100713a.html
5. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association : JAMIA. 2010 Mar-Apr;17(2):131-5.
6. McCarty CA, Wilke RA. Biobanking and pharmacogenomics. Pharmacogenomics. 2010;11(5):637-41.
7. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. The American Journal of Human Genetics. 2010;86(4):560-72.
8. Pace WD, Cifuentes M, Valuck RJ, Staton EW, Brandt EC, West DR. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009;151(5):338.
9. South BR, Shen S, Jones M, et al. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics. 2009;10(Suppl 9):S12.
10. Grishman R, Huttunen S, Yangarber R. Information extraction for enhanced access to disease outbreak reports. J Biomed Inform. 2002;35(4):236-46.
11. Xu H, Jiang M, Oetjens M, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. Journal of the American Medical Informatics Association. 2011;18(4):387-91.
12. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000:270-4.
13. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
14. Chapman W, Bridewell W, Hanbury P, Cooper G, Buchanan B. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001;34(5):301-10.

15. Ananthakrishnan AN, Cai T, Savova G, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflammatory bowel diseases. 2013;19(7):1411-20.
16. Ananthakrishnan A, Gainer V, Perez R, et al. Psychiatric co-morbidity is associated with increased risk of surgery in Crohn's disease. Alimentary pharmacology & therapeutics. 2013;37(4):445-54.
17. Savova GK, Olson JE, Murphy SP, et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. Journal of the American Medical Informatics Association. 2012;19(e1):e83-e9.
18. Lin C, Karlson EW, Canhao H, et al. Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records. PLoS One. 2013;8(8):e69932.
19. Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. Methods Inf Med. 1998 Nov;37(4-5):334-44.
20. Friedman C. A broad-coverage natural language processing system. Proceedings / AMIA Annual Symposium AMIA Symposium. 2000:270-4.
21. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. May 1;17(3):229-36.
22. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A, 3rd. The KnowledgeMap project: development of a concept-based medical school curriculum database. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2003:195-9.
23. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13.
24. Apache cTAKES. http://ctakes.apache.org/. 2013
25. Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the I2B2 project. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:931.
26. Murff HJ, FitzHenry F, Matheny ME, et al. Automated Identification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing. JAMA: The Journal of the American Medical Association. 2011;306(8):848-55.
27. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010;26(9):1205.
28. Xu H, Jiang M, Oetjens M, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. Journal of the American Medical Informatics Association : JAMIA. 2011 Jul-Aug;18(4):387-91.
29. Wagholikar KB, MacLaughlin KL, Henry MR, et al. Clinical decision support with automated text processing for cervical cancer screening. Journal of the American Medical Informatics Association. 2012;19(5):833-9.
30. Vleck TV, Elhadad N. Corpus-based problem selection for EHR note summarization. In: AMIA annu symp proc; 2010. p. 817–21.
31. Cunningham DH, Maynard DD, Bontcheva DK, Tablan MV. GATE: A framework and graphical development environment for robust NLP tools and applications. 2002.
32. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering. 2004;10(3-4):327-48.
33. Kano Y, Baumgartner WA, McCrohon L, et al. U-Compare: share and compare text mining tools with UIMA. Bioinformatics. 2009;25(15):1997.
34. Hahn U, Buyko E, Landefeld R, et al. An overview of JCoRe, the JULIE lab UIMA component repository. 2008; 2008. p. 1–7.
35. v3NLP. https://wiki.chpc.utah.edu/display/htx/v3NLP+Framework+Tool+Development (accessed 1 Oct 2013). 2013
36. Coden A, Savova G, Sominsky I, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform. 2009;42(5):937-49.
37. Penn. http://repository.upenn.edu/cis_reports/570/. 2013
38. Wu ST, Kaggal VC, Dligach D, et al. A common type system for clinical natural language processing. J Biomed Semantics. 2013;4(1):1.
39. ORBIT. National Library of Medicine. ORBIT: Online Registry of Biomedical Informatics Tools. 2011. http://orbit.nlm.nih.gov 2013

40. NLP Ecosystem. http://nlp-ecosystem.ucsd.edu/events/webinars/making-natural-language-processing-nlp-more-accessible-analysis-clinical-text. 2012
41. Apache. http://incubator.apache.org/guides/proposal.html. 2013