Huntsman Cancer Institute Optimizes Research With Linguamatics NLP Platform

The Research Informatics Shared Resource (RISR) is the research informatics team at the Huntsman Cancer Institute (HCI), University of Utah. The RISR team is tasked with capturing and reporting patient information to improve the quality of data capture, and to make it more accessible for research initiatives. Because of the vast amount of critical patient information stored in unstructured formats, finding specific data is often challenging. After years of compiling information manually or using rudimentary natural language processing (NLP) tools, seven years ago HCI Research Informatics adopted the Linguamatics NLP platform. Today HCI has more efficient research processes, allowing it to collect data faster and provide its researchers with higher quality data.

QUICK FACTS

Situation: HCI at the University of Utah is a nationally recognized cancer center that relies heavily on data for its research studies. Accessing high quality data was often time consuming, because so much information was stored as unstructured text within clinical records. RISR developed robust and effective manual and automated processes to capture data into disease-specific research repositories. However, there still was an abundance of data that existed in unstructured and structured documents such as surgical pathology reports and physician clinical notes.

Solution: HCI implemented the Linguamatics NLP platform to develop NLP rules, which has enabled more efficient data capture and more measurable quality information. After a successful proof-of-concept project working with breast cancer data, HCI expanded to other areas, including studies into prostate and hematologic malignancies.

Success: Using the NLP platform, HCI can automatically extract data that is of high research quality. This has helped HCI to improve its research efforts and enhance disease understanding in support of achieving better outcomes. HCI presented a poster at the 2018 American Society of Hematology (ASH) national meeting in San Diego on its work on myeloid neoplasms. HCI's Samir Courdy, in a long-standing collaboration with City of Hope, is also authoring a joint paper to submit for publication on non-Hodgkin's lymphoma.

Situation

HCI at the University of Utah is a nationally recognized cancer hospital and research center in Salt Lake City. HCI's RISR team supports HCI by developing comprehensive computing and information systems that allow researchers to accomplish their studies more efficiently.

Samir Courdy is the director of RISR and Chief Research Information Officer for HCI. His team is responsible for developing tools and systems to capture data that advances quality of care efforts and enables research. "We capture data in all forms, whether it is genomic, clinical, bio-specimen or population-based studies," explained Courdy. "In addition to collecting patient data for quality of care initiatives, we improve access to data for research purposes so that users have high quality data for studies or to apply for research grants."

Gaining access to high quality data, however, can be a challenge. While finding demographic data or coded disease information is relatively straightforward, researchers also need to access critical details such as morphology, topography, tumor size and stage. Often these insights are not easily extracted and remain trapped in narrative-style surgical or clinical notes, or in pathology and radiology reports.

Manual abstraction of data is very time consuming and can be prone to errors. It may take a user several hours to track down, review and transcribe the critical elements from individual patient records, and because documents must be individually reviewed, even a diligent researcher may miss critical details.

Over the years, Courdy explored various technologies to improve the search process and automate the capture of high quality data. "About 10+ years ago we began developing internal methods for using publicly available natural language processing tools," said Courdy. "We used the technology to automate breast cancer research, but felt the process was too lengthy and tedious, and wasn't yielding results at the scale we were hoping for. It was also fragile and did not translate well to other disease areas."

Courdy continued looking for solutions and eventually encountered Linguamatics at an Informatics conference in 2011. "After learning about the Linguamatics NLP tools and reviewing the [...] solution, we decided to abandon the old technology we had been using for NLP and started using [the Linguamatics NLP platform]. And the rest is history."

Solution

Once HCI implemented the NLP platform, HCI's first project focused on breast cancer. "We started with breast cancer because we had already built a collection of pathology reports with patient cohort data and were able to use the same gold standard for the [...] toolset," said Courdy. "Initially we trained [the NLP platform] on the breast cancer data set while working closely with the folks at Linguamatics. This project was intended as a proof-of-concept for using [the NLP platform]—and the concept worked."

Following the success of the initial project, the RISR team expanded into other disease areas. "For example, we built a very robust process for prostate research and then focused on hematologic malignancies, starting with non-Hodgkin's lymphoma," said Courdy. Since first launching the Linguamatics NLP platform, HCI has expanded its use and developed NLP tools for multiple other conditions.

> ***"Many times, we need data that is stored as unstructured text and not easy to find. Looking for information manually is a very costly endeavor, but that is the way we did it for many years."***
>
> *— Samir Courdy, Director of RISR and Chief Research Information Officer, HCI*

Courdy noted that HCI's work with the NLP platform and Linguamatics has opened the doors for collaboration with other cancer institutes. "In 2014, I went to a big data conference and was talking about the work we were doing with [the NLP platform]," said Courdy. "Someone in the audience asked if we would be willing to share certain queries. Linguamatics agreed to explore the idea and today all our queries are shareable, which benefits everyone."

In fact, once Linguamatics helped enable the sharing of NLP rules, HCI began collaborating with other institutions. "We were able to work with another group on a non-Hodgkin's lymphoma study and now, after sharing data back and forth, we are finalizing a paper for publication," said Courdy.

Success

HCI's use of the NLP platform has made its collection processes faster and more efficient, and provided researchers with higher quality data to advance research initiatives. Two specific areas of improvement include:

- **Faster data capture:** Courdy noted that, "Now we can capture more data faster on many, many more documents than an individual could do manually. I don't have a quantifiable number, but we significantly improved our processes and the quality and the access to the data."

- **Higher quality data:** HCI's use of NLP has improved access to higher quality data, which in turn has advanced research efforts and enhanced disease understanding in support of achieving better outcomes. Since adding the Linguamatics NLP platform, HCI has successfully published several studies and presented research findings on hematology malignancies at the 2018 ASH conference.

"We're able to provide our principal investigators with the quality data they need to apply for or qualify for grants, write papers or identify cohorts for specific studies," said Courdy. "For example, when applying for grants you have to prove that you have the depth and breadth of data to support your research. With [the NLP platform] we are able to demonstrate that we have the tools needed to facilitate access to high quality data."

For other organizations implementing the NLP platform to advance research efforts, Courdy offered the following advice:

1. Start by identifying the specific problem you are trying to solve. Once the problem is well defined, you will have a better understanding of the data you are looking for, allowing you to narrow your field of search.
2. Find the right people to perform the work and provide them with training opportunities. Give them time to learn the internal workings of the NLP platform so that they understand how best to utilize it in a very efficient way. Allow them to attend formal training sessions and work on a small prototype project.
3. Have a team to provide support for users on the NLP platform framework. Verify that data within an electronic warehouse can be accessed electronically, and is available in a format that the NLP platform can read and rules can be applied.

"Research organizations like ours, that have massive amounts of data, need technology that automates the capture of relevant data," said Courdy. "This facilitates the delivery of high quality data that is relevant; provides insights into patients' treatments, outcomes and lifestyle habits; guides the development of therapies; and provides researchers with the information they need for grant applications and additional studies."

"Linguamatics [NLP platform] is the driver for collecting high quality data and making the process more efficient."

— Samir Courdy, Director of RISR and Chief Research Information Officer, HCI

CONTACT US
+44 (0)1223 651 910 (U.K.) | +1 617 674 3256 (U.S.)
nlp@iqvia.com
linguamatics.com