Swansea University
Prifysgol Abertawe

College of Medicine, Swansea University

Accelerating medical research and improving patient care with new insight into healthcare records

Overview

The need

With data on patients, diseases and treatments locked away in unstructured text records, Swansea University was concerned that the lack of insight was holding back important medical research work.

The solution

The University will be uncovering new insights hidden across billions of anonymized healthcare records, using a first-of-its-kind content management and analytics platform from IBM and Open Connections.

The benefit

The solution will open up a rich set of healthcare data that can help organizations make better decisions about treatments and patient care and deliver faster answers to research questions, improving public health.

Every day, vast quantities of structured and unstructured content are collected about patients as they pass through health service organizations. This wealth of information holds valuable insight that can be used to drive research breakthroughs and improve care. However, with content growing faster than most organizations can consume it, unlocking the power of this information is often a struggle.

The Health Informatics Group at Swansea University's College of Medicine is working to develop a powerful platform for analyzing vast quantities of unstructured content, built on IBM Watson™ Foundations software.

For the first time, researchers will have access to a vast collection of information that can be easily mined for new insights, helping researchers find answers to complex questions and supporting health systems in driving higher quality care.

"Merging structured and unstructured data provides a more complete picture of patients and conditions that can be used to improve the way we treat diseases and deliver care," says Professor David Ford, Professor of Health Informatics and Director of Health Informatics Group, College of Medicine, Swansea University.

Solution components

Software

- IBM® Content Analytics
- IBM Content Classification
- IBM Content Collector for File Systems
- IBM FileNet® Content Manager
- IBM InfoSphere® Guardium® Data Redaction

IBM Business Partner

- Open Connections

Transforming billions of records into new insight

Since 2006, the Health Informatics Group has been working with NHS Wales and local and central governments to bring together routinely collected patient-centric electronic data in a central resource known as the Secure Anonymized Information Linkage (SAIL) Databank.

To date, the group has loaded more than nine billion records from hundreds of health and social care service providers into the SAIL databank, through a carefully designed de-identification and privacy protection system. Sources include regular extracts of all hospital inpatient stays, day cases and outpatient visits, along with data from hundreds of Welsh GP practices and other health and non-health service providers.

While initial work has focused on processing structured data, the Health Informatics Group is now seeking to tap into the potential of the huge volumes of unstructured data held in SAIL.

Professor David Ford, Professor of Health Informatics and Director of Health Informatics Group at Swansea University's College of Medicine, elaborates: "Unstructured content has always been the most prevalent way of collecting and storing information in the healthcare sector – around 80 percent of medical data is unstructured, found in documents such as electronic medical records (EMRs), physician notes, medical correspondence and more.

"Traditionally, it has been incredibly difficult to analyze this kind of content on a large scale. As content analysis relies on the ability to detect patterns and relationships, when data is not consistent and neatly structured it is much more challenging to identify and extract concepts of interest.

"To make the SAIL databank a truly exceptional resource for our research and service provider partners, we need an easy, accurate way of finding concepts of interests in a blizzard of free text. As we are dealing with billions of records, querying data manually is out of the question – we need the help of some very clever computing."

As most of the information held in SAIL relates to patients and healthcare organizations, the Health Informatics Group faces the additional challenge of ensuring that the identity of individuals is safeguarded at all times, as Professor David Ford explains: "We must comply with very strict data protection legislation and confidentiality guidelines, and make sure that any details that could potentially be used to identify an individual are eliminated before data is loaded into SAIL."

"The solution from IBM and Open Connections is opening up a very exciting future for researchers and service providers, giving us access to an enormous and rich data source that we really had no way of analyzing before."

— Professor David Ford, Professor of Health Informatics and Director of Health Informatics Group, College of Medicine, Swansea University

Forming a close partnership

The Health Informatics Group is working closely with a number of partners to develop a first-of-its-kind platform for mining vast quantities of de-identified unstructured content. At the heart of the new platform is the Open Connections Text Analytics Solution (OCTAS), built on leading-edge technology from IBM and Open Connections, an IBM Business Partner.

"We have a strong pre-existing partnership with IBM, who have provided much of the core infrastructure for the SAIL databank," notes Professor David Ford. "Now that we are making new inroads into working with unstructured data, IBM technology is once again proving to be a valuable resource.

"Similarly, the decision to partner with Open Connections was an easy one. They are experts in IBM Enterprise Content Management solutions and they have proven expertise in the healthcare space. We felt confident that Open Connections could work with us to create a tailor-made solution for processing and analyzing data."

De-identifying data

OCTAS is based on much of the same technology used by IBM Watson. The Health Informatics Group uses IBM Content Collector for Files to collect unstructured content from hundreds of sources. Anonymized records are automatically categorized and organized with IBM Content Classification software, and stored in a central IBM FileNet® Content Manager repository. This unstructured content is then analyzed using the intelligent text mining and natural language processing capabilities offered by IBM Content Analytics.

Professor David Ford describes how data is de-identified: "SAIL is based on a privacy protection approach called the 'separation principle'. We provide tools that help our data providers divide information into two parts – one containing purely medical content, such as diagnosis and treatment information, with no personal identifiers, which is sent directly to SAIL.

"The data provider then takes the remaining set of data, containing only identifying information including a patient's name, address, date of birth and, where applicable, NHS number, but no sensitive medical information and submits it electronically to a secure Trusted Third Party system operated within the NHS. This generates a unique but meaningless identifier for each individual. The identifier is then sent to SAIL, where it is recombined with the rest of the content and then fully encrypted.

"Content that once sat unused will be made available to analytics, helping uncover patterns of cause and effect and indicators of disease that were previously unknown."

— Professor David Ford, Professor of Health Informatics and Director of Health Informatics Group, College of Medicine, Swansea University

"As a result, we can link data that was previously static and siloed in different systems to create a full record of a patient's treatment lifecycle – one that can be easily updated with new information as it is received by SAIL. And we can do all this in a totally secure, anonymous way, ensuring that citizen identities remain completely protected."

He continues: "Our current challenge, something we are actively working on with Open Connections, is to use the IBM toolset to automatically seek and reliably redact personal identifiers within free text documents whilst they are still within the NHS, and do so in a way that will automatically feed into our existing anonymization and linkage system. We will then be able to load the de-identified content into the SAIL Databank, and link the new records to the existing content, without increasing the privacy risk of the data. This will massively increase the potency of SAIL for a huge number of research uses."

Finding meaning in masses of information

Once a robust methodology for de-identifying data has been developed, the Health Informatics Group will be able to extract meaningful information from the vast store of free text held in SAIL, using content analytics and natural language processing tools.

"Essentially, we will be 'training' the platform to process language like people do, examining the structure and words in a document to derive meaning from it," comments Professor David Ford. "Once the text has been processed, the next step is to use content analytics to extract concepts of interest, such as social history, lifestyle factors, support available at home, functional ability and outcomes, as well as their relationships to each other.

"Combining this knowledge with structured data will enable us to discover patterns and relationships across all the records held in SAIL, providing a more complete picture of patients and treatment patterns or the spread of diseases. The ability to process and link data in this way can help provide faster answers to very complex and precise research questions – something that no one physician could discover without many months or years of research."

"The potential for advancing research and improving patient care is huge – and we are excited to work with our partners to shape this new generation of information analysis and discovery."

— Professor David Ford, Professor of Health Informatics and Director of Health Informatics Group, College of Medicine, Swansea University

Delivering higher-quality healthcare

Professor David Ford continues: "One area where we anticipate the new platform will make a big difference is around the patient's personal circumstances, preferences and history. When a doctor first sees a patient, he or she typically records a patient history, to understand an individual's lifestyle, symptoms, previous treatments and care preferences. As this information is recorded as free text, it has been largely inaccessible to researchers up until now.

"As we train the platform to automatically draw out key themes scattered across unstructured data, it will open up a whole new set of information to researchers, so they can find answers to questions such as, 'How much does poverty affect the need and demand for health services?', 'How much improvement in patient outcomes is being achieved by a particular treatment or intervention?', or 'Are there enough patients suitable for a new clinical trial in a specific area in Wales?' Armed with these insights, healthcare organizations can better understand the core issues affecting patient care and make more informed decisions, helping shape a more effective, patient-centric health system."

Supporting vital research

Beyond building a better healthcare experience, the new platform is creating new opportunities for researchers. The Health Informatics Group is already working on early pilot projects with a number of research teams in the fields of epilepsy, multiple sclerosis and cardiovascular disease, who will be using the platform to help uncover new patterns of disease that can be used to improve treatment and even save lives.

"The solution from IBM and Open Connections is opening up a very exciting future for researchers and service providers, giving us access to an enormous and rich data source that we really had no way of analyzing before," remarks Professor David Ford.

"Our researchers will be able look at data on millions of individuals and test their hypotheses on a huge sample size – effectively, treating the whole of Wales as a laboratory – without entailing the costs of collecting new data. Content that once sat unused will be made available to analytics, helping uncover patterns of cause and effect and indicators of disease that were previously unknown. The potential for advancing research and improving patient care is huge – and we are excited to work with our partners to shape this new generation of information analysis and discovery."

About Swansea University

Founded in 1920, Swansea University is a public, research-led university located in Swansea, Wales. With some 14,500 undergraduate and postgraduate students and more than 2,500 staff, the University produces internationally excellent and world-leading research across all disciplines.

For more information, please visit: www.swansea.ac.uk

About the Health Informatics Group

The Health Informatics Group is part of the College of Medicine at Swansea University. The main aim of the group is to realize the potential of electronically held, person-based, routinely collected information for the purposes of conducting and supporting health-related research. Among many things, the group develops and operates the SAIL databank, is part of the MRC-funded Farr Institute, and is an ESRC-funded Administrative Data Research Centre.

To find out more, please visit: www.healthinformatics.swansea.ac.uk

About Open Connections

As one of the longest established IBM partners for the FileNet product range, Open Connections has a huge depth of experience helping customers build enterprise-class content management and business process management solutions that enable organizations to securely capture, store, retrieve, manage and analyze content.

To learn more about products, solutions and services from Open Connections, please visit: www.openc.co.uk

OPEN CONNECTIONS

For more information

To learn more about IBM Enterprise Content Management solutions, contact your IBM representative or IBM Business Partner, or visit the following website: **ibm.com**/software/ecm

IBM

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
August 2014

WVC12354-USEN-01