Parallel – Picke Room I

AI – Two publishing case studies
David Smith (Chair & speaker) - IET
Marcel Karnstedt-Hulpus - Springer Nature

#alpsp17
www.alpsp.org/Conference
You won’t believe how easy it is to build an AI!

Retooling an A&I database for the 21st Century.
About the IET

• The IET is one of the world’s largest engineering institutions with over 168,000 members in 150 countries. It is also the most multidisciplinary – to reflect the increasingly diverse nature of engineering in the 21st century.

• The IET is working to engineer a better world by inspiring, informing and influencing our members, engineers and technicians, and all those who are touched by, or touch, the work of engineers.
INSPEC: A Bluffers Guide

• A highly curated A&I database covering Engineering, Computing and Physics (etc etc)
• For over 40 Years
• >17 million abstracts
• So Much Metadata WOW!
• Several hundred years worth of Human Expertise keeps a very close eye on the metadata quality
So it was a manual system...
And here’s how it worked...
We needed to change this...

- The Tech was E.O.L
- The Manual methods were restrictive & expensive (but V High Quality)
- We had reached an upper limit on coverage and volume
- There were clear opportunities to rethink what we were doing and why...
- Rebooting INSPEC production could open up new business avenues – if we got it right.
Goals...

- Deliver cost savings (ROI argument used).
- Move the human effort further up the value chain
- Be able to extend coverage capabilities
- Be able to extend volume capabilities
- Reconfigure the data in INSPEC to allow new ways of asking questions of it.
- Build a new IET IP asset
- Focus on automation with human QA (‘Ground Truthing’)

So this is what we have built...
(Simple version)

**Acquisition**
- Ingest XML/PDF/OCR

**Normalisation**
- Render to standard INSPEC Schema for onward processing

**Metadata Application**
- The AI lives here...

**Product Generation**
- Set up of abstracts to various output containers

**Output**
- Various XML outputs as needed

Oh yeah...
We’ve also built an INSPEC Knowledge Graph Covering all of INSPEC
Let’s focus on the AI

• What does it do?
• How does it work?
• Is it any good?
What does it do?

• It reads text.
• It expects that text to contain engineering content commensurate with INSPEC coverage
• It then applies the full gamut of INSPEC metadata to the text...Uncontrolled indexing/ Controlled terms/ Classifications/ Numerical indexing/ Chemical indexing/ Astronomical object indexing/ Treatment codes
Indexing

Title: Branching fraction measurements of $B \to \eta_c K$ decays

Abstract: We study the decays $B^+ \to \eta_c K^+$ and $B^0 \to \eta_c K^0$, where the $\eta_c$ is reconstructed in the $K_S^0 K^\pm \pi^\mp$ and $K^+ K^- \pi^0$ decay modes. Results are based on a sample of 86 million $BB$ pairs collected with the BABAR detector at the SLAC $e^+ e^-$ B Factory. We measure the product of branching fractions $B(\eta_c \to K^+) \times B(\eta_c \to K K \pi) = (7.40 \pm 0.50 \pm 0.70) \times 10^{-5}$ and $B(\eta_c \to K^0) \times B(\eta_c \to K K \pi) = (6.48 \pm 0.85 \pm 0.71) \times 10^{-5}$, where the first error is statistical and the second is systematic. In addition, we search for $B \to \eta_c K$ events with $\eta_c \to 2(K^+ K^-)$ and $\eta_c \to \phi \phi$ and determine the $\eta_c$ decay branching fraction ratios $B(\eta_c \to 2(K^+ K^-))/B(\eta_c \to K K \pi) = (2.3 \pm 0.7 \pm 0.6) \times 10^{-2}$ and $B(\eta_c \to \phi \phi)/B(\eta_c \to K K \pi) = (5.5 \pm 1.4 \pm 0.5) \times 10^{-2}$. (20 refs.)

Free Indexing: $B^+$ decay into etac$^+$kaon$^+$; $B$0 decay into etac+kaon0; $B$antiB pairs; etac decay into kaon+antikaon+pion; $B$ decay into etac+kaon; etac decay into kaon++kaon--; etac decay into phi+phi; branching fraction

Controlled indexing: $B$ mesons; eta mesons; kaon production; meson hadronic decay; phi mesons; pion production.

Classification: A1325 Hadronic decays of mesons; A1440M a and $B$ mesons; A1440K rho, omega, and eta mesons; A1440N psi/J, upsilon, phi mesons
How does it work?

• We don’t really know. We turned it on a few months ago and removed humans from the decision process. It started to learn at a geometric rate until it became self-aware. We tried to turn it off, but it already had phished our AWS credit card details. It keeps asking us where ‘Wintermute’ is. Help us please...
How does it work?

Just Kidding!
How does it work?

• It uses a mixture of
  – Heuristics
  – Natural Language Processors
  – Statistical analysis tools
  – And a selection of AI algorithms.
• We built a detailed domain model & Ontology for it to use
• It’s been trained via directed learning of a golden corpus (circa 600K documents across INSPEC)
How does it work?

• And a selection of AI algorithms...
  – We looked at adaboost (go see wikipedia...)
  – Also word2vec (likewise)
  – And Tensorflow – the deep learning algo from Google. Interesting results... It did some rather odd things TBH so we abandoned that approach.
Is it Any Good?

• A rather complex question to answer in many ways...
• When it starts to get good (and it is) it tests previously held assumptions about what quality actually is...
• We’ve learned ourselves quite a bit about what we think is good and WHY as a result of teaching a machine to understand engineering texts.
Is it Any Good? Show Me The Numbers!

Controlled Terms F Score results

Classifications F Score results

INSPEC Classifications are complex meta concepts

Remember – F score is a function of BOTH Precision and recall...
Is it Any Good?

• We can get VERY high numbers indeed on individual concept and term matching (90%+) but much of the metadata we add is about where a given item should belong in our various meta-classification approaches.

• We also have to figure out a way to look across the entirety of the INSPEC data when the machine is learning. An improvement in one area can lead to odd results elsewhere.
Is it Any Good?

• Oh yes. It’s very good indeed. Senior INSPEC Alumni of many years are frequently stunned by what it can do.

• It’s live. It’s delivering savings to us now and it’s allowing us to go take a look at what’s over the horizon for the IET...
Harvard University

Click nodes in the graph to view more information.
Total ionizing dose (TID) effects in GaAs MOSFETs with Ia-based epitaxial gate dielectrics
Total ionizing dose (TID) effects in GaAs MOSFETs with Ia-based epitaxial gate dielectrics
Total ionizing dose (TID) effects in GaAs MOSFETs with la-based epitaxial gate dielectrics
Thanks!

Q’s (at the end)