Building an AI-enabled open-access research tool: The Insight Engine 2.0

John Herr
Duke University Office of Information Technology; Software Engineer, Durham, NC, USA; john.herr@duke.edu

The first version of the Insight Engine (I_E) was built in 2010, and was envisioned by William Seaman (Duke University) and designed and developed by Todd Berreth (now North Carolina State University) and Oliver Perriquet, computational linguist, (then) Le Fresnois, France. The system brought together a linked transdisciplinary Search Engine, Database and Visualization systems to promote intelligent information generation and flows. More recent work was contributed by Matthew Kenney (now working with Apple). The application was written in PHP and later Ruby on Rails and visualizations were made using the Processing and P5.js libraries. In the early version intelligence in the system was provided by word2vec and doc2vec models generated from Gensim NLP pipelines, and by SQL queries, all of which operated on a large corpus of research documents. The first version of I_E was a closed system; participation was by invitation only, and the intelligence in the system was limited to the information held in the corpus of documents and artifacts and intelligent information processing. Insight Engine 2.0 is an open system which anyone will be able to use. It will enable research teams to interact and contribute papers and other annotated media elements. The system is strengthened with information coming from more diverse scholarly sources such as journal articles published online in addition to the information held in the corpus of research that was already in the database of the first version of Insight Engine. Still lead by Seaman, the design and development of I_E 2.0 is performed by John Herr, Software Engineer (Duke Office of Information Technology), Hojung (Ashley) Kwon, Computer Scientist (Duke ‘21), Dev Seth, Computer Scientist (Duke ‘22), and Quran Karriem, PhD candidate, Computational Media, Arts and Cultures, Duke University, and Mingyong Cheng, PhD candidate at UC San Diego.

The goal of the Insight Engine 2.0 is to build an information–oriented digital world for human and machine users (“Micropeers”) to interact within. That goal leads us to consider both the human user experience (UX) and machine user experience (MX) of our system, and after consideration of both needs we chose the Unity gaming engine as the foundation for our informatin-driven world. Unity gives us contexts of physics, 3-d space, and time, allowing us to relate bodies to one another through mathematical relationships that can be expressed as distance and as forces of attraction or repulsion. We are approaching the following set of information-related goals — We seek to define a series of Information Processing Structures that function on multiple levels of data mining, date exploration and subsequent use by interactant researchers; we seek to map aspects of Sentience production and re-embody them in a biomimetic model system that will contribute to making a prototype for Neosentience (Seaman’s coin) production in an intelligent robotic system; we seek to design resilient and robust interactive information processing systems which will intelligently help to articulate relationalities between disparate kinds of research discourse by providing links and bridges to key concepts that might be lost in a more traditional search setting.

One way to achieve these goals is to reflect the actions of humans and Micropeers (collecting, highlighting, archiving, and performing work on the artifacts in the system) in the environment via stylistic and dynamic design changes to the interface of the I_E. The world of the Insight Engine can become a mood or an aesthetic which reflects the nature of the interactions being performed in many ways through the articulation of specific choices by the interactant, the artifacts that are being related to,

the information those artifacts contain, and the insight our system can potentially generate from them. Additionally, core concepts such as “benevolence” inform the presentation of the world through intelligent queries which affect global attributes including color, typography, and shape, and other aspects like gravity and/or attraction to facilitate users’ interaction with the Insight Engine. The refinement of those styles can reflect not only the UX needs, but also the MX needs, and could lead to attunement “pathways” (the building of intelligent bridges across disciplinary boundaries) which could be visualized to human users in order to yield insight through and into machinic intelligence. Other emergent patterns of visualized text and data will yield insights with the aim of sparking new ideas and transdisciplinary research approaches. The larger design aesthetic for I_E 2.0 incorporates metaphors from nature and biology, including molecules, forests of information, and planets, giving us multiple contexts to show microscopic and macroscopic growth and evolution in relation to space and time as they exist between bodies, objects, and ideas. The Unity engine provides a real-time style transfer to facilitate these features, making it possible for I_E to appear in different ways to different users in surprising and unexpected ways which are highly pointed in terms of potential informational growth.

In order for I_E to be truly open it must itself be open to information from outside systems. In general there will be two levels of users – the general public, and alternately researchers who will actively contribute links and relevant papers and media objects. A foundation of knowledge and factual information about the physical world must be present in order for the system to generate meaningful affordances that can be grasped by both human and information-oriented machine bodies. This brings into the play the 4Es of Cognition — Embodied, Embedded, Enactive, and Extended approaches to knowledge production. We chose the Diffbot natural language knowledge graph API as the means by which our system would gain information and relationships about the world and, internally, in relation to our collected corpus. Other collections of textual works (and/or annotated media objects like diagrams, scans, 3d Models, etc.) can also be curated into I_E or be added by an invited user of the system to become part of the database. Our system also uses machine learning pipelines and AI queries via Python Transformers to gain clarity and insight in relation to the specific research which is being performed. Systems with dependencies are susceptible to change from the outside, and so as a practical matter we have attempted to mitigate unintended effects from without by putting middle-ware layers in-between Diffbot and our Unity front-end, giving us the ability to change knowledge graph providers in the future, or to create our own graphs. A choice of front-end technology is a major commitment, and we feel confident that Unity (and React, which we also use) will stand a reasonable test of time while being well-supported into the future by their respective authors. Given that this will be an open source system, we imagine unique new functionalities in the future. Unity is also incredibly performant and can render on a variety of devices including phones, desktop browsers, and virtual headsets, making it the most resilient of choices. Practically speaking, we use dockerized containers deployed to Kubernetes clusters and other flexible infrastructures. This approach also eases on-boarding new developers or potentially opening our development up to the open source community in the future in a programmatic, information-relevant manner.

We look forward to sharing more details and experiences resulting from the rebuilding of the Insight Engine.