EVC Enterprise Ventures Corporation

Data Collection, Extraction, Conversion

CTC Enterprise Venture Corporation's (EVC's) Open Source Intelligence (OSINT) and collection experts have the ability to identify and collect structured, semi-structured, and unstructured data from the Internet to meet a client's project requirements. Our team is currently collecting and converting semi-structured data such as business directories, normalizing, and converting it to the client's format at an average rate of 12,000 records per hour, based on the quality of the source data. In addition, our team has developed a custom extraction tool for unstructured records. Utilizing a Human-in-the-Loop process to verify fields and content, our approach can be rapidly modified to adapt to changes in the data format or content.

Data Collection and Conversion

Utilizing proven screen scraping technologies, EVC's data conversion team builds customized strategies for collecting data from electronic resources. Once collected, the data is analyzed for format and quality to identify steps to convert the data into the required format. This includes normalizing phone numbers and addresses to the client's requirements, extracting elements such as email address or phone number from fields such as "Comments" to populate appropriate fields, or building relationships between elements such as "subsidiary" or "owned by." This process can be applied to legacy files to prepare historical data for new database formats or improved data mining activities.

Unstructured Data Extraction

Frequently data may be contained in documents not suitable for Natural Language Processing (NLP), such as documents with sections in upper-case letters, or limited punctuation or grammar. Our Human-in-the-Loop process can handle a wide variety of document contents and formats. Our customized extraction scripts ensure that the elements critical to the client are identified and appropriate extraction routines are developed. Once elements are initially tagged, our data conversion team reviews for accuracy of field content or tagging. For example, is "Denver" the name of a city or the name of a person? Our team reviews context in order to tag and format the data appropriately. The results are formatted to feed into other analytical tools, such as Palantir, or into an existing database.

For more information, contact:

Gregory Jablunovsky
Director, Professional Services Program Development
814.262.6497
gregory.jablunovsky@evc.ctc.com

Jeffrey Anderson
Managing Director, Professional Services
814.262.6867
jeffrey.anderson@evc.ctc.com

EVC Enterprise Ventures Corporation
Innovative Solutions for a Complex World

www.evc.ctc.com