Smartlogic™

FACT SHEET

Semaphore Fact Extraction Framework (FACTS)

Fact Extraction Framework - FACTS

Semaphore's Fact Extraction Framework (FACTS) brings together key Semaphore technologies that provide Information Scientists, stakeholders and subject-matter experts access to a powerful, intuitive and user-friendly fact extraction and structured classification process.

FACTS bridges the gap between a businesses' understanding of its information and content, and the desire to extract facts from that content without having to understand the configuration, semantic rules or NLP capabilities.

The FACTS framework provides:

- → A methodology for fact extraction and structured classification that exploits the rule-based linguistic power inherent in Semaphore
- → A user-friendly, browser-based, graphical UI-that supports the methodology
- → A publisher configuration that auto-generates rules to streamline and simplify implementation

FACTS uses Semaphore's native semantic rules and NLP processing techniques to generate higher-order rules that transform unstructured and semi-structured text into highly structured, concrete data elements that can be used in downstream business application processing (i.e. graph database, BI and RPA, etc.).

What is a Fact?

Facts - units of information contained within content that have a specific meaning - can have a simple (person, organization, date, or predefined concept from a taxonomy) or complex structure.

Complex facts allow for the creation of multiple user-defined facts. For example, an Address fact, where street address elements are grouped separately from person elements. When combined, both facts form a composite Address fact, as shown in Figure 1.

- Address fact
  - Person identify fact
    - Person SSN
    - Person name
      - Person first name
      - Person family name
      - Person salutation
  - Street Address fact
    - (Apartment number – optional)
    - Building number
    - Street name
    - Town / city
    - ZIP
    - State

Figure 1. Address fact, which contains 2 smaller facts; Person identify fact and Street Address fact.

FACTS can be configured to return:

- Machine-learning trained entities such as:
  - People, locations, dates, currency, addresses, organizations, products, email addresses, URIs, etc.
  - FACTS has access to the full Semaphore entity recognition set in English and other languages
- Model-driven concepts (with GUIDs / URIs) using any taxonomy in Semaphore
- The raw text of any part of a document; sentences, paragraphs, user-defined range of words such as legal clauses, regulatory definitions, etc.

FACTS Methodology

When a user understands their content - the documents, their structure, and the data and facts they contain, they can model specific aspects of that content to allow them to extract relevant facts.

Smartlogic uses a structured and well-tested methodology to extract facts from documents:

1. Identify the **facts** you want to extract
2. Identify the **content** that contains the facts
3. Create a document **fingerprint** that identifies the document type using appropriate context
4. Create **extractors** to extract the facts using specialized context and fact types

The FACTS methodology is flexible and supports many use cases and domains.

User-friendly UI to Support the Methodology

Semaphore Knowledge Model Management (KMM) is a web-based, feature-rich platform used by model builders to collaborate with stakeholders and subject matter experts (SME) to create models that accurately reflect the organization. Our easy-to-use interface is designed for business users to support model building processes without lengthy training.

The FACTS framework is a Semaphore KMM licensed plugin. When the FACTS Framework model is visible in the KMM workspace you can link it, as well as other required models (i.e. domain taxonomy/ontology), into your project model.

The FACTS model uses intuitive and easy-to-understand constraints to visually guide the user to create valid FACTS models by showing them the optional and mandatory properties.

FACTS modelling supports a basic framework to represent the different elements of the extraction methodology as outlined below:

- Document Type - Allows you to identify different document types / templates found within your corpus that have a material impact on fact extraction. For example, only certain facts are found in certain document types.
- Document Metadata - Identifies the universal fact for any document type. All document facts are extracted by contexts.
- Context – Defines context based on fact patterns found within your content. Contexts can contain other contexts and allow you to build complex grouped user-defined facts.
- Facts – Allow for the definition of different fact types, i.e. facts driven by a taxonomy, entity zone, or wildcard / regex patterns. Fact types are used to perform textual matching - finding the text we are interested in.

Where the Magic Happens – Publisher

The FACTS framework uses the model-driven, rules-based, and linguistic capabilities of Semaphore to:

- Generate the rules modelled in KMM by publishing the FACTS project model
- Leverage the rules in the classification process to extract the relevant facts
- Process the content using the generated facts and Classification and Language Services

The publishing process generates the appropriate rules to make the extractors and facts work using a common sense approach and addresses the idiosyncrasies of how rules work within Semaphore to achieve the desired result.

Fact Extraction Rules and Classification

The FACTS framework allows organizations to identify structured and contextualised facts from documents, which are used in further processing by machine learning, AI, and business intelligence (BI) systems. The systems can then focus on the discovery of latent patterns and knowledge not previously known to your business.

Conclusion

Semaphore’s FACTS framework is a conceptual model-driven approach to representing business content. It bridges the gap between SME’s, who understand their content and the Information technical specialists, who understand the process of extracting it.

With FACTS, organizations can leverage the value of all information – structured and unstructured – to drive key business initiatives.