In a Nutshell: a 2-phase Frame-based Approach

Phase 1 – Linguistic Feature Extraction
By performing several standard NLP tasks, a mention-based structured representation of the input text is built, organizing all the annotations produced by NLP tools (e.g., NERC, EL, TERN, SRL) in an RDF graph of mentions (i.e., spans of text denoting some entities or facts).

Phase 2 – Knowledge Distillation
The mention graph is processed via SPARQL rules to distill a knowledge graph, where each node uniquely identifies an entity of the world, event or situation, and arcs represent relations between them (e.g., the participation and role of an entity in an event).

RDF Data Model for Information Extraction

Graphical Rendering of Extracted Knowledge

SPARQL-based Knowledge Distillation
Various types of SPARQL rules: instance creation, typing, naming, DBpedia linking, frame-role linking, coreference resolution

Example (Instance Creation for Argument Nominalization):

```
INSERT { ?m ks:denotes ?i ; ks:simplies ?f ; ks:expresses ?g .
WHERE { ?m a ks:FrameMention ; nif:anchorOf ?a ; ks:roleset ?s .
?i a ks:Instance .
?f a ks:Frame .
?a a ns:me.mentionOf .
?g a ns:ex.mentionOf .
} } .
```

Post-processing: inference, smushing, redundancy elimination, compaction.

Performances

Detecting and representing frames and frame-role relations:
→ precision: 0.716
→ recall: 0.494

Processing large document corpora (Simple English Wikipedia):
→ 110K pages in about 507 core hours
→ processing-time linearly scales with the size of the text
→ 0.85 accuracy in extracting triples about DBpedia entities

References:
- Corcoglioti, F., Rosspocher, M., Palmero Aprosio, A., Extracting Knowledge from Text with PIKES. In ISWC Posters & Demonstrations, 2015.