The European Commission's science and knowledge service

Joint Research Centre

Directorate I - Competences

JRC I.3 Text and Data Mining Unit

European Commission

Timelines

Jakub Piskorski, Vanni Zavarella, Martin Atkinson

Text and Data Mining Unit
Directorate I: Competencies
Joint Research Centre of the European Commission
Ispra, Italy

11 April 2019

Timelines

- **Goal:** build a core engine able to **acquire**, **scan** and **analyse** large text collections of various types in order to **compute time-ordered series of open-domain events involving the target entity** of the search
- **Motivation:** law-enforcement, tax, customs, migration and other agencies perform **similar-in-nature entity-centric searches in large document collections (web)** in order to extract structured information on the target entity for facilitating higher-level analysis
- **Research area:** Open Information Extraction, Temporal Reference Extraction and Reasoning, Named-Entity Extraction, Visual Analytics

Philosophy

A complex system that works is invariably found to have evolved from a simple system that worked. **A complex system designed from scratch never works and cannot be patched up to make it work.** *You have to start over with a working simple system.*

John Gall (1975)

Development criteria

- trade-off between 'state-of-the-art' academic results and scalability
- time and space efficiency
- configurability
- multilinguality

Timelines

Search Query:

NAME: Donald Trump
GENDER: male
TYPE: person

Variants:
Trump
D. Trump
Donald John Trump
Trump's

Trump was born on June 14, 1946, in Jamaica, Queens, a neighbourhood in New York City.

→ EVENT TYPE: **BIRTH**
EVENT PHRASE: **was born**
TIME: June 14, 1946
RELATED ENTITIES: Jamaica, Queens, New York City

Trump attended Fordham University in the Bronx for two years, beginning in August 1964.

→ EVENT TYPE: **PARTICIPATION**
EVENT PHRASE: **attended**
TIME: August 1964
RELATED ENTITIES: Fordham University, Bronx

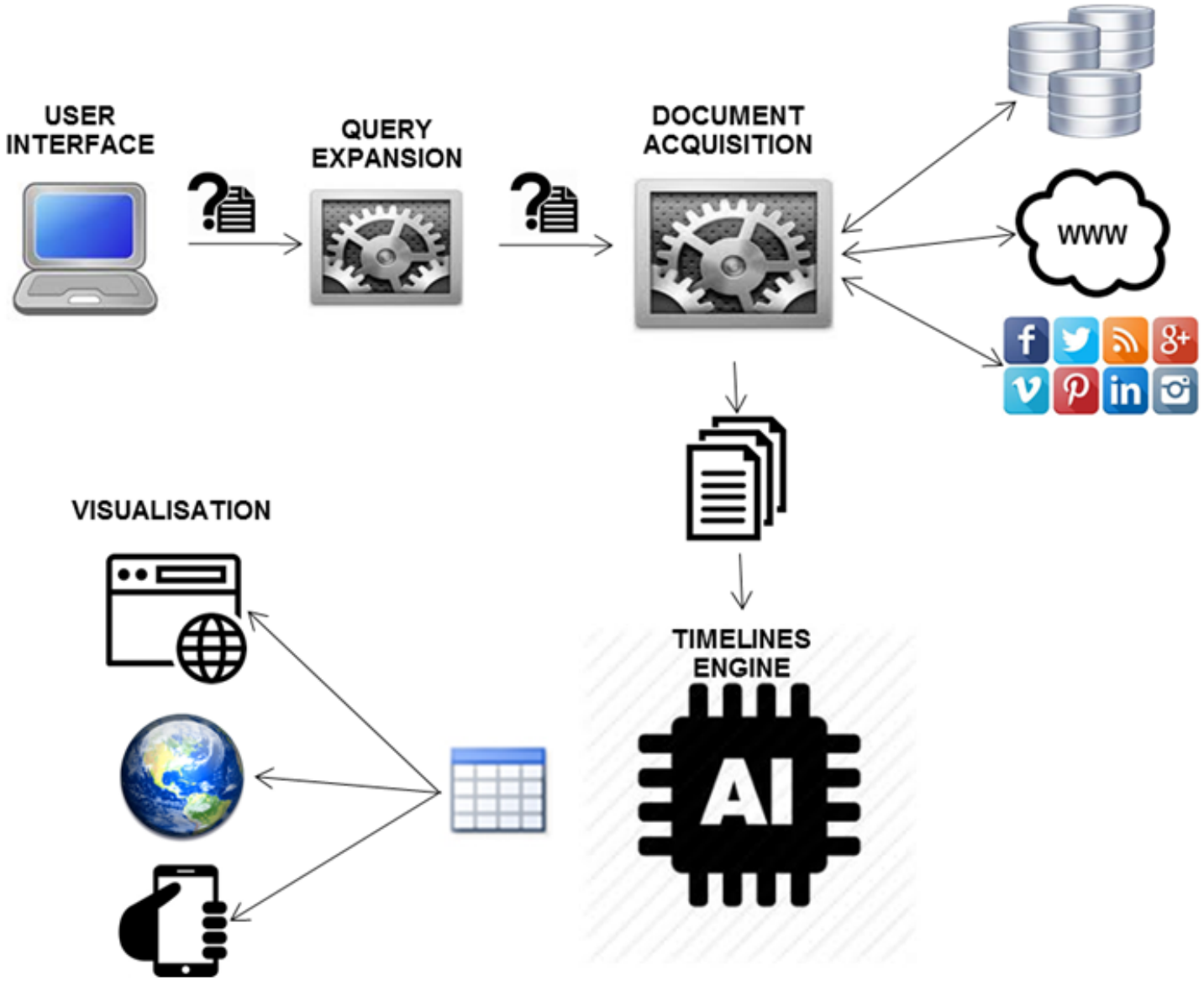
While there, from 1968 he **worked at** the family's company, **Elizabeth Trump & Son**, named for his paternal grandmother.

→ EVENT TYPE: **WORK-FOR**
EVENT PHRASE: **worked at**
TIME: 1968
RELATED ENTITIES: ...

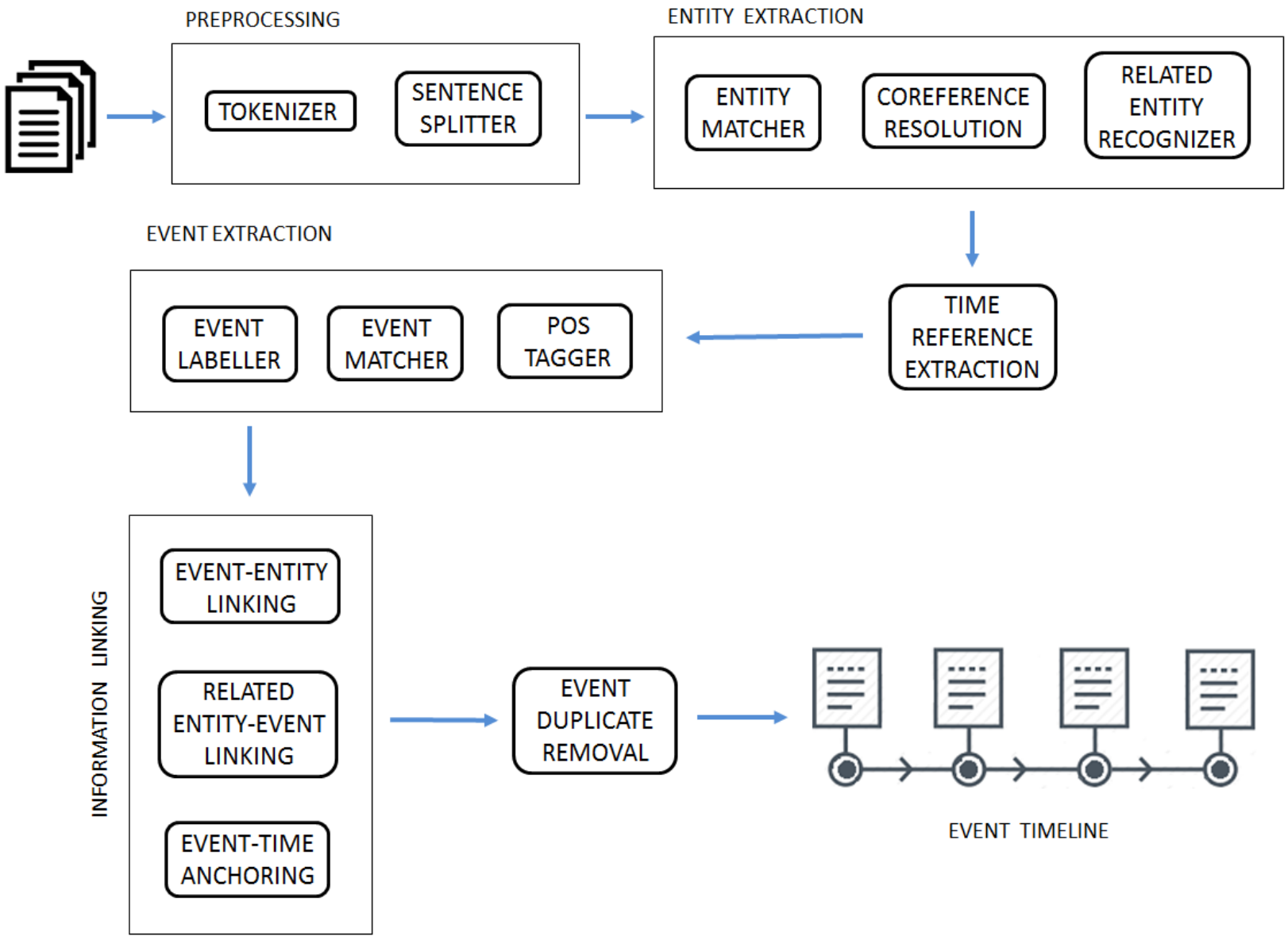
Since 1971, he **has chaired** **The Trump Organization**, the principal holding company for his real estate ventures and other business interests.

→ EVENT TYPE: **MANAGEMENT**
EVENT PHRASE: **has chaired**
TIME: Since 1971
RELATED ENTITIES: ...

Timelines: Big Picture



Timelines: Architecture



Timelines: Processing Resources

- **Pre-processing:** in-house CORLEONE toolkit
- **Entity Extraction:**
  - Target Entity Matching:
    - exact match ('*Trump*')
    - expansion ('*Trump Foundation*')
    - (parameterizable) fuzzy match ('*Trump's*')
  - Coreference Resolution:
    - only pronominal anaphora (e.g. '*he*', '*her*') within sentence
  - Related Entities:
    - exploit ca **13 mln** NE lexical resources: JRC Variant Names, BabelNet and GeoNames
    - guessing rules based on capitalisation
- **Time Reference Extraction:** in-house FSS-Timex module

- **Event Extraction:**
  - POS Tagging: Stanford POS Tagger
  - Event Matcher: finite-state grammars for recognition of Lightweight Verb Constructions (e.g., '*has met with*') and Nominalisations (e.g., '*Attendance*')
  - Event Labelling: lexical rules for tagging with coarse-grained categories (OWNERSHIP, LAW-RELATED, STATEMENT, ...)
- **Information linking:**
  - Target Entity-to-Event Linking: heuristics
  - Event-Time Anchoring: heuristics (document context)
  - Related-Entity-to-Event Linking: all in the sentence
- **Event Duplicate Removal:** text (near-)duplicates

Example

Preprocessing	*Donald Trump made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed he signed the waiver Wednesday.*
Entity Extraction	**Donald Trump** **[1]** *made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed* **he** **[1]** *signed the waiver Wednesday.*
Time Reference Extraction	**Donald Trump** **[1]** *made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed* **he** **[1]** *signed the waiver* **Wednesday (29-11-2017)** *.*
Event Extraction	**Donald Trump** **[1]** *made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed* **he** **[1]** **signed** *the waiver* **Wednesday (29-11-2017)** *.*
Information Linking	**Donald Trump** **[1]** *made no reference to signing a waiver that officially delays any move of the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed* **he** **[1]** **signed** **[1]** *the waiver* **Wednesday(29-11-2017)** **[1]** *.*

Output (JSON)


```
{ "eventMatch":
    { "docID": 16462,
      "creationDate": "7-12-2017",
      "startPosition": 6,
      "endPosition": 25,
      "startPositionInDoc": 4248,
      "endPositionInDoc": 4267,
      "confidence": 1.0,
      "eventStartDate": "29-11-2017"
      "eventEndDate": "29-11-2017"
      "textSnippet": "Trump made no reference to signing a waiver that  officially delays any move of
                      the U.S. Embassy from Tel Aviv to Jerusalem, but the White House confirmed he
                      signed the waiver Wednesday."
      "eventDescription": "signed"
      "eventCategory": [ "CREATION" ]
      "targetEntityParticipation": true
      "nominalized": false
      "coordinated": false
      "eventType": []
      "tense": "PAST"
      "aspect": "UNSPECIFIED"
      "modality": "null"
      "relatedEntities": [ { "entity":"U.S. Embassy", "cat":"OTH", "matchType":"EXACT-NAMED-MATCH", },
                           { "entity":"Tel Aviv", "cat":"OTH", "matchType":"EXACT-NAMED-MATCH" },
                           { "entity":"Jerusalem", "cat":"OTH", "matchType":"EXACT-NAMED-MATCH", },
                           { "entity":"White House", "cat":"OTH", "matchType":"EXACT-NAMED-MATCH", } ]
    }
  "targetEntityMatches": [ { "form":"Trump", "matchType":"EXACT-NAMED-MATCH" },
                           { "form":"he(Donald Trump)", "matchType":"PRONOMINAL" } ]
}
```

Accuracy (ball park estimates)

100 Events with "mention" of **Donald Trump** randomly selected from ca. 60K events extracted from 100K news article corpus.

category	performance
Event detection accuracy (all)	97%
(only factual)	81%
Fraction of events assigned at least one category	70%
Event categorisation accuracy	91%
Target-Entity-Event-Linking accuracy (strict)	76%
(lenient)	74%
Related Entity Extraction precision	96%
recall	91%
Event-Time Anchoring coverage	33%
precision	63%

Efficiency

Computing timeline for **Donald Trump** on 100K news articles:

- **Number of events detected:** 154049
- **Number of Events after removing duplicate:** 68514
- **Total processing time:** 370 seconds

tokenization	0.5\%
sentence splitting	1.4\%
transformations	0.2\%
timex resolution	27.3\%
pos tagging	7.6\%
entity matching	16.6\%
related entity recognition	7.9\%
event detection	9.8\%
time-event anchoring	0.1\%
entity-event linking	0.1\%
related entity-event linking	0.0\%
event category labeling	0.1\%
anaphora resolution	0.1\%
logging	0.1\%
timeline sorting	0.0\%
duplicate removal	0.2\%
other	27.8\%

Timelines: Future

- general accuracy improvement
- linking/grouping events using semantic text similarity metrics
- multi-linguality
- on-demand user-based event classification
- full-fledged application

Timelines: Demo

https://test.emm4u.eu/Timelines/