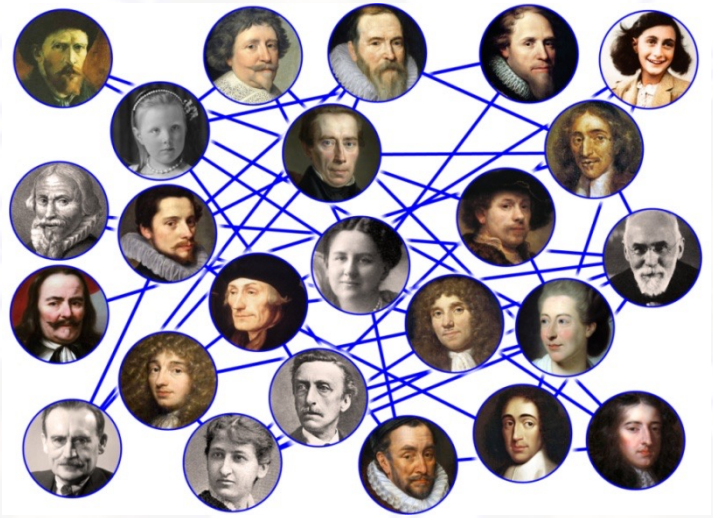
BiographyNet

Linking the world of History



Workshop on Biographical Linked Data

Friday 22 January 2016

Team BiographyNet (http://www.biographynet.nl)

netherlands eScience center

huygens ing

VU UNIVERSITY AMSTERDAM

The beginning

- 23 biographical resources
- Covers ± 80,000 people in ± 145,000 biographies
- Biographical text and various metadata
- Found at: http://www.biografischportaal.nl

Guus Schreiber

Els Kloek

Susan Legêne

Piek Vossen

Main project theme

What kind of historical questions can be answered with this data with the help of computational methods?

Interdisciplinary team

Niels Ockeloen

Computer Scientist

Serge ter Braake

Historian

Antske Fokkens

Computational linguist

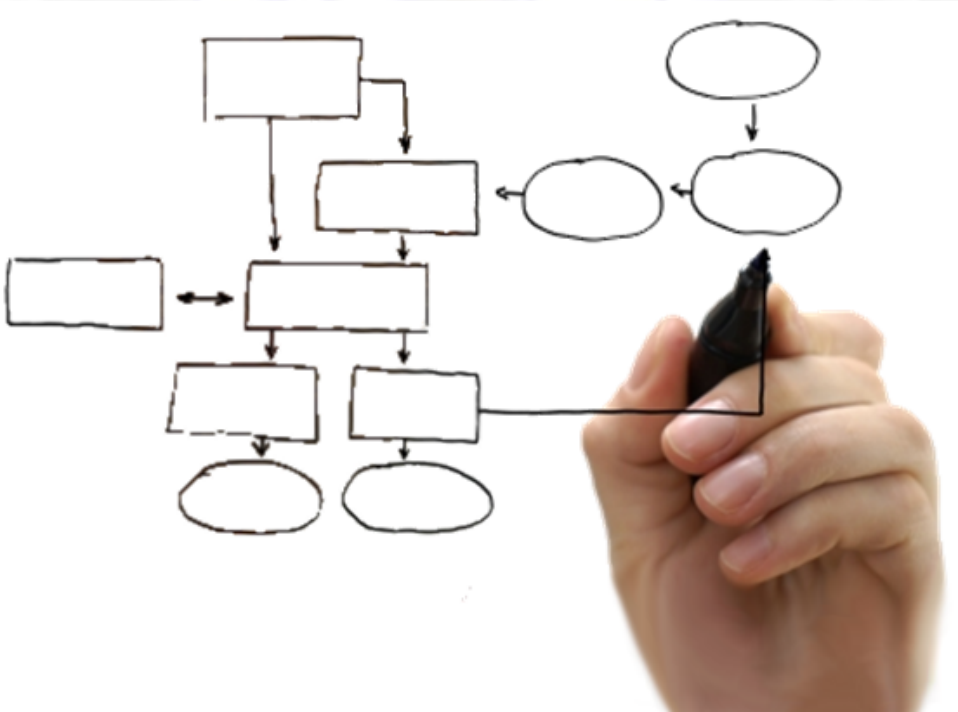
Historian's questions

- Occurrences of concepts & people
- Group analyses:
  - educational background
  - age when obtaining function
- Overall corpus statistics:
  - men versus women
  - Horoscope of people
  - Focus on specific century

About the Data

RDF version of the original 'Biografisch Portaal' data

- Schema based on the structure of the original XML files
  - Needs to facilitate the coupling of different biographies of the same person, without compromising the original data
  - Compatible with existing schemas such as EDM, PROV, P-PLAN, DC terms, etc.
- Some numbers about the original data:
  - 8,014,356 triples
  - 327.869 places (mentions)
  - 315,500 events
  - 110,648 biographies
  - 76,359 persons
  - 54.395 dates
- SPARQL endpoint at: http://data.biographynet.nl

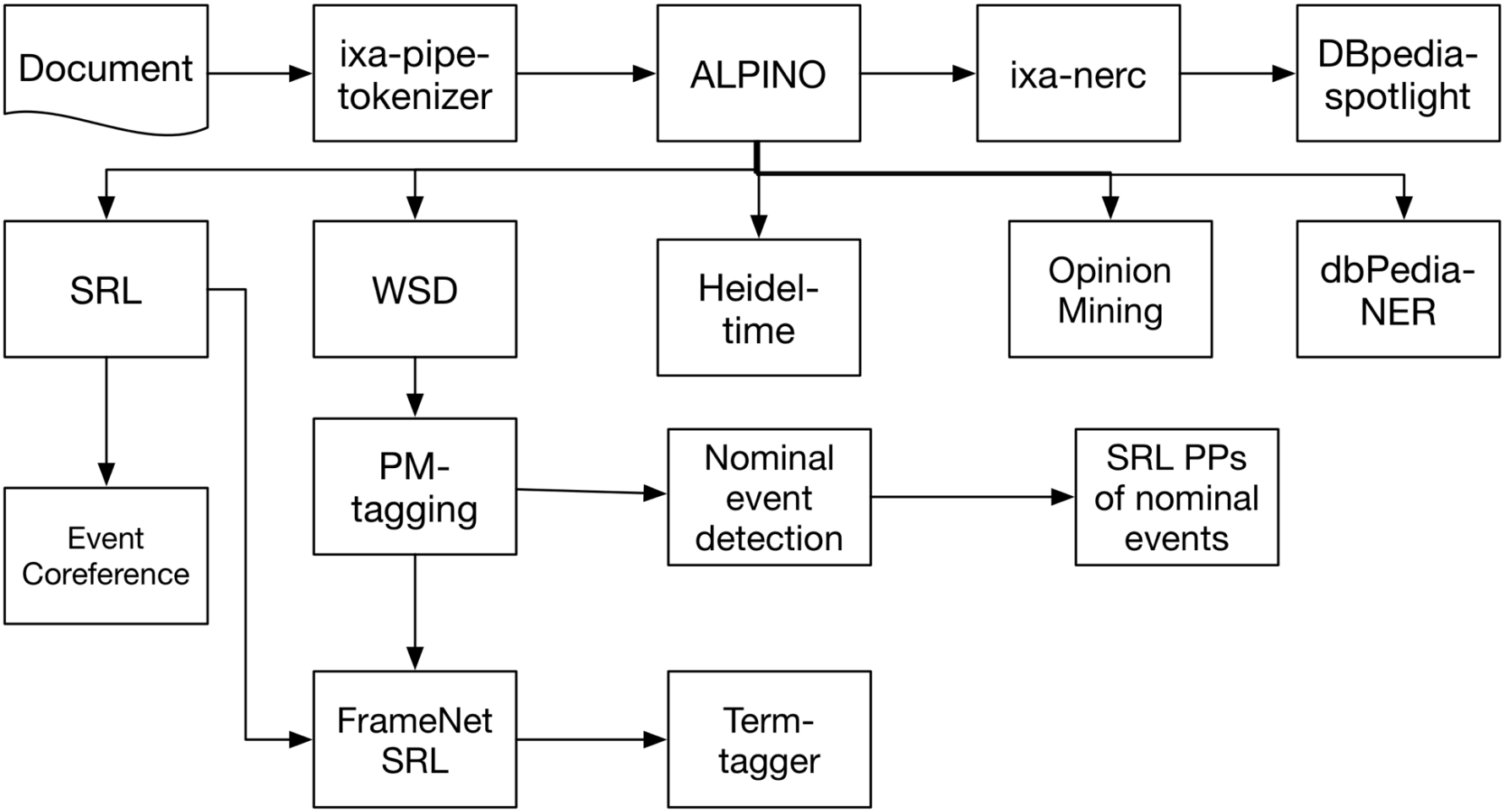


netherlands eScience center

huygens ing

VU UNIVERSITY AMSTERDAM

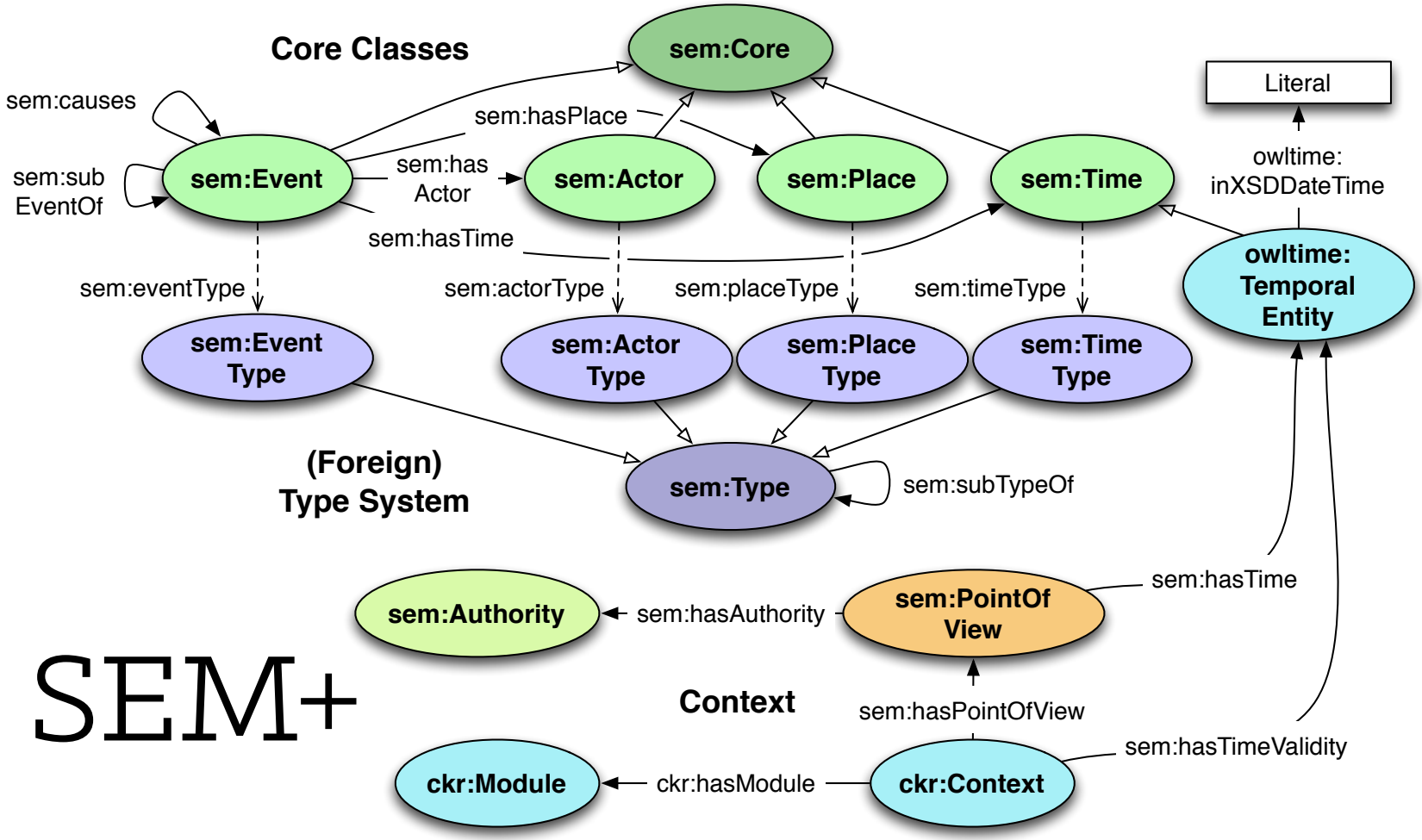
Dutch pipeline



Interpretation

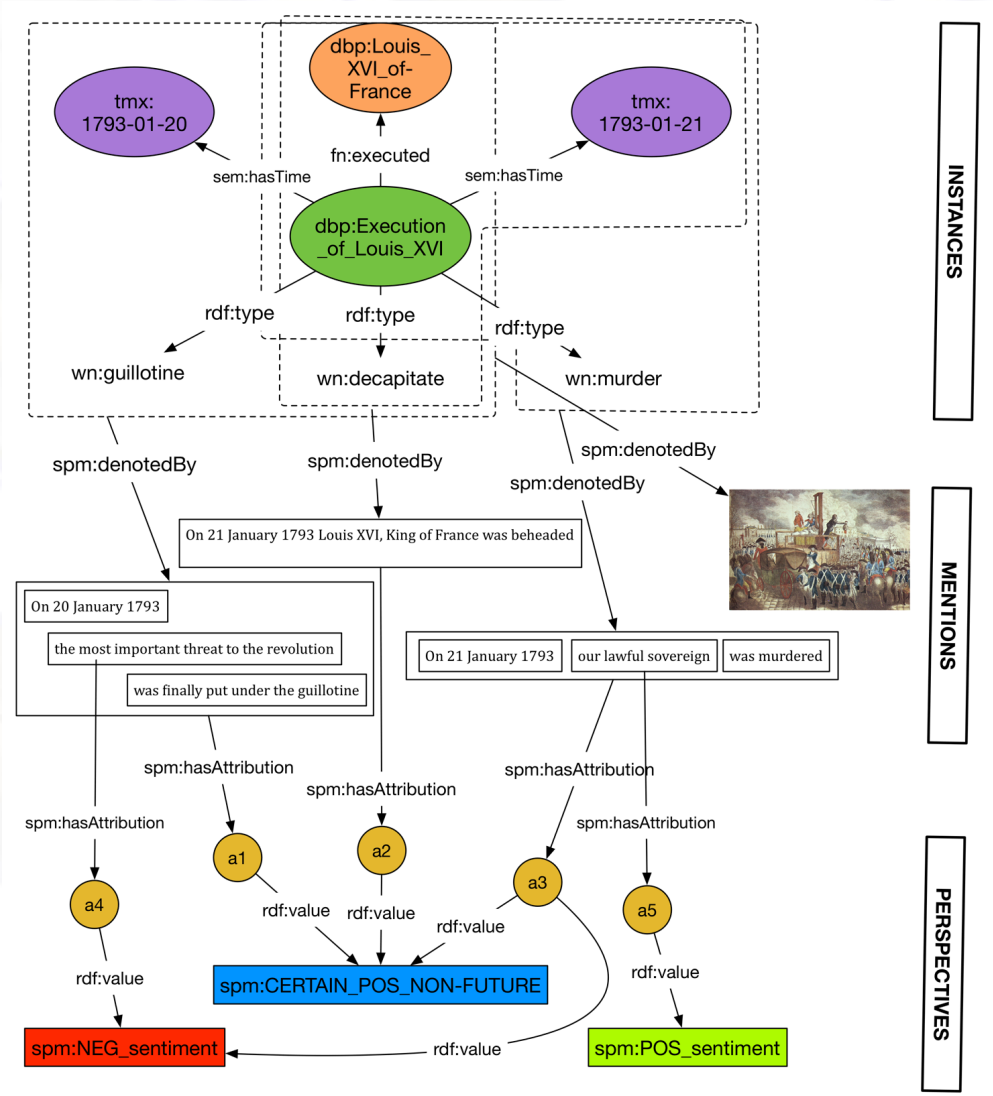
- Translate NLP output to RDF:
  - Simple Event Model
  - Grounded Annotation Framework
  - BiographyNet schema
- Targeted interpretation for highly relevant information:
  - Core events
  - Family relations
  - Whose profession?

SEM+

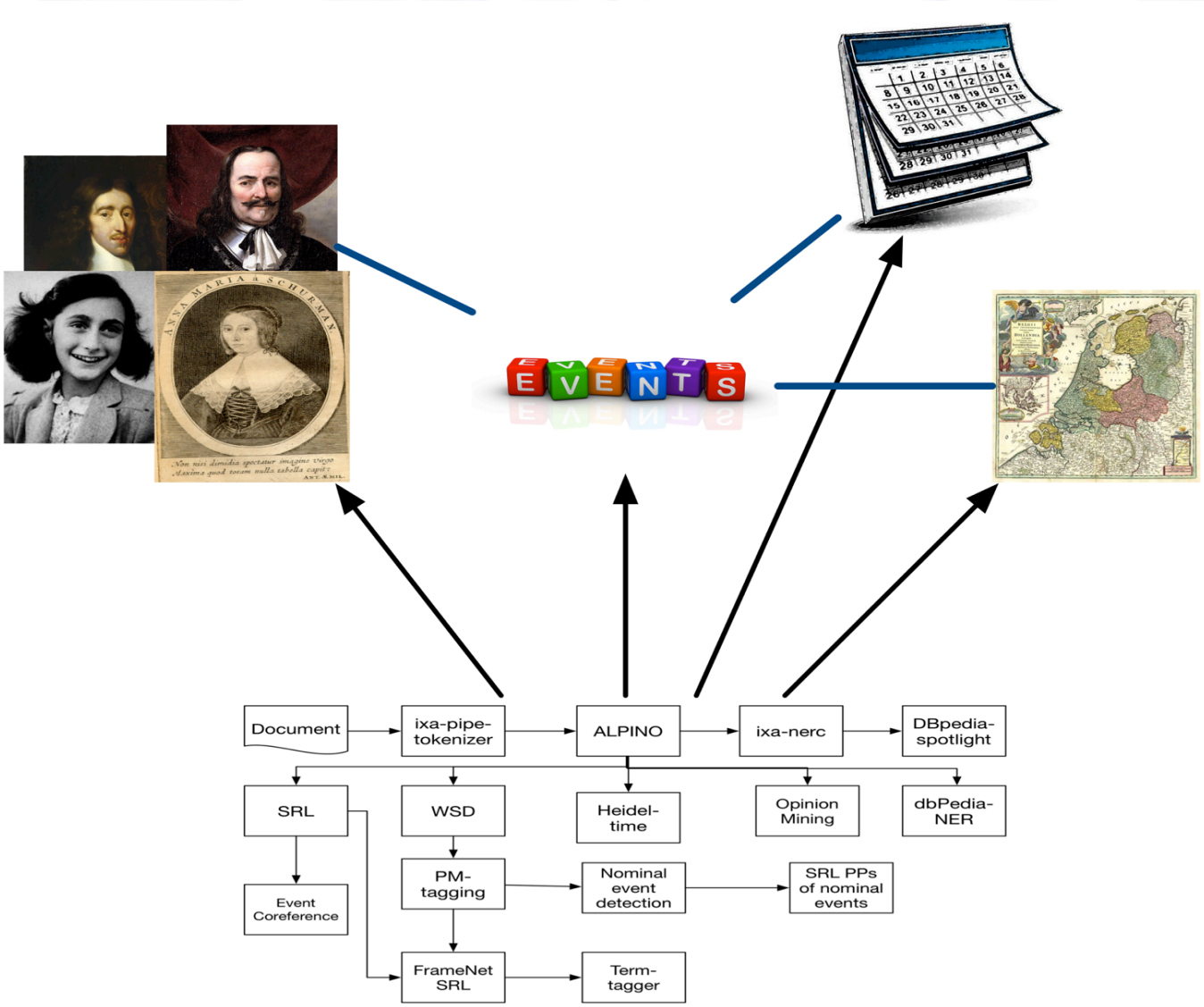

Core Classes
sem:Core
Literal
sem:causes
sem:hasPlace
sem:sub
EventOf
sem:Event
sem:has
Actor
sem:Actor
sem:Place
sem:Time
owltime:
inXSDDateTime
sem:hasTime
owltime:
Temporal
Entity
sem:eventType
sem:actorType
sem:placeType
sem:timeType
sem:Event
Type
sem:Actor
Type
sem:Place
Type
sem:Time
Type
(Foreign)
Type System
sem:Type
sem:subTypeOf
sem:hasTime
sem:Authority
sem:hasAuthority
sem:PointOf
View
Context
sem:hasPointOfView
sem:hasTimeValidity
ckr:Module
ckr:hasModule
ckr:Context

SEM+

GAF (and GRASP)



Pipeline Output



Interpretation output I

Huismeester

Dichteres

fn:Education_teaching

fn:Education_teaching@subject

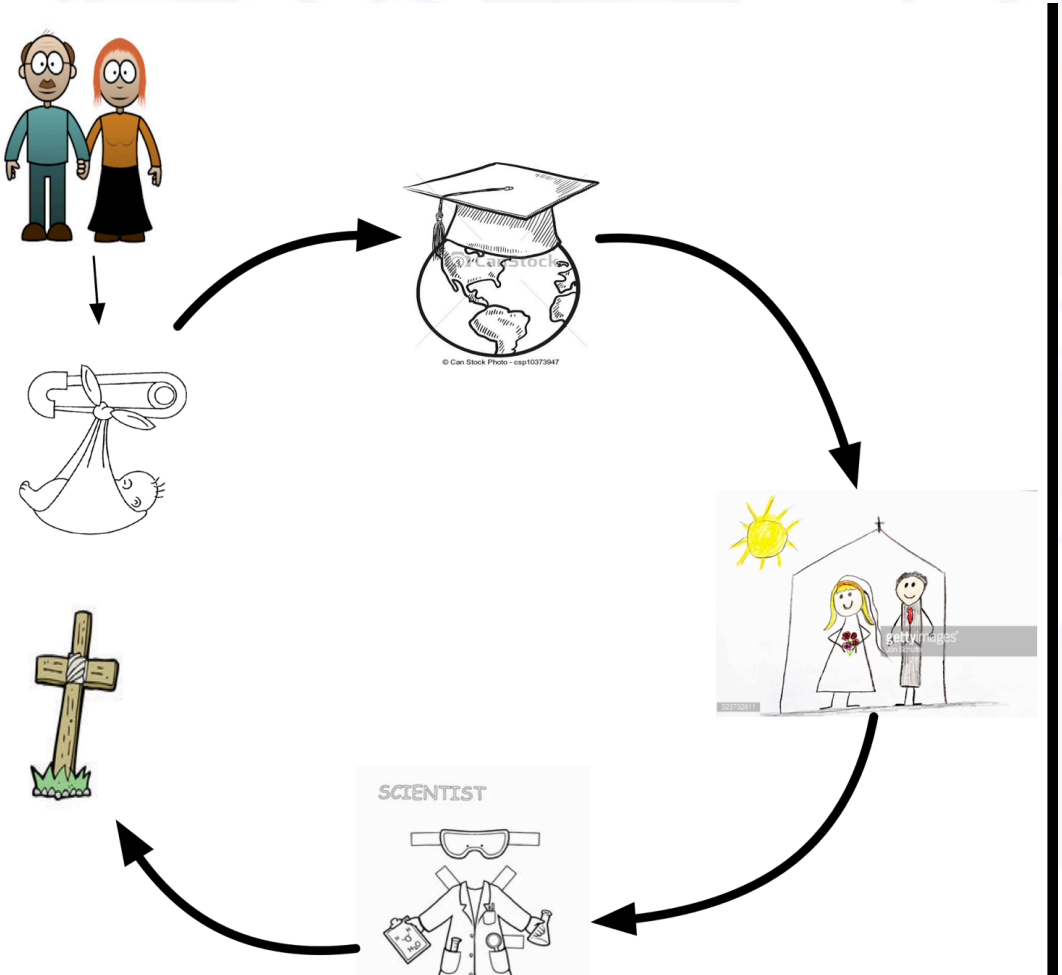
WordNet:ili-30-02547586-v

fn:Education_teaching@student

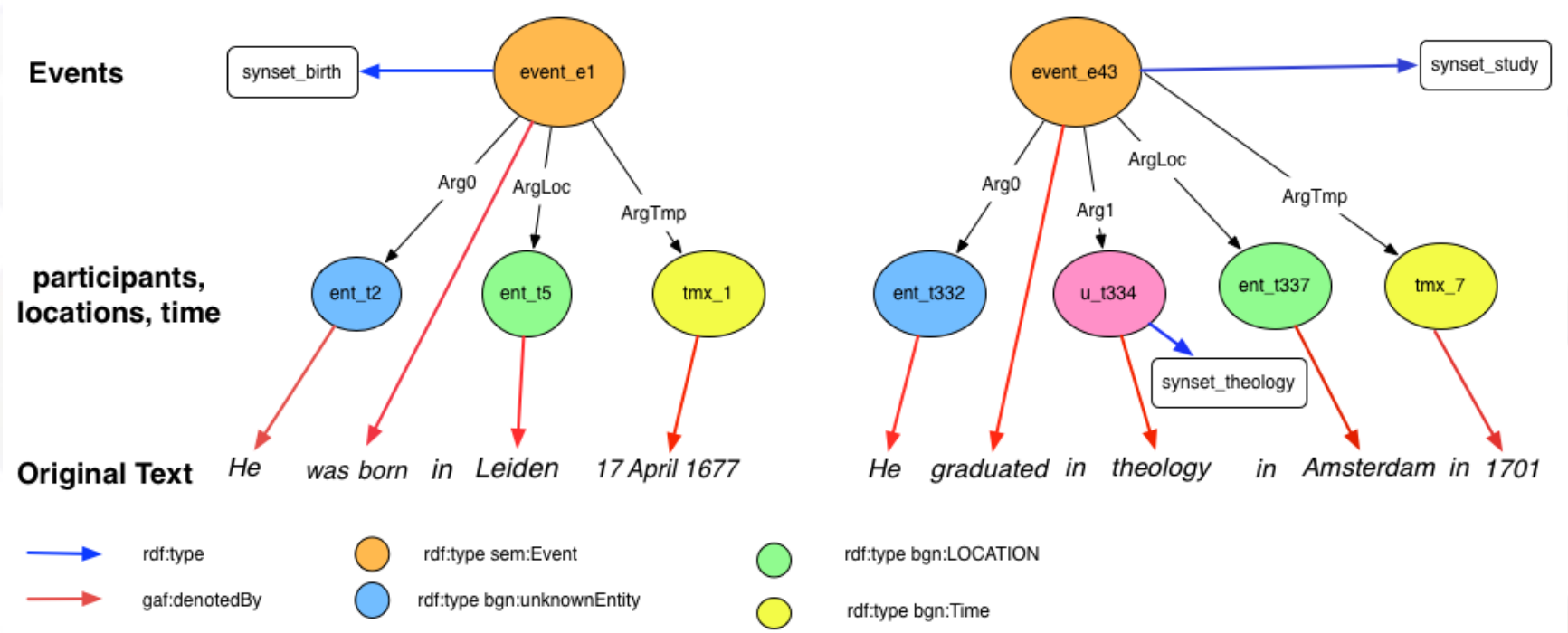
“theologie”

pron_fsg

Output Interpretation



Event example (detail)



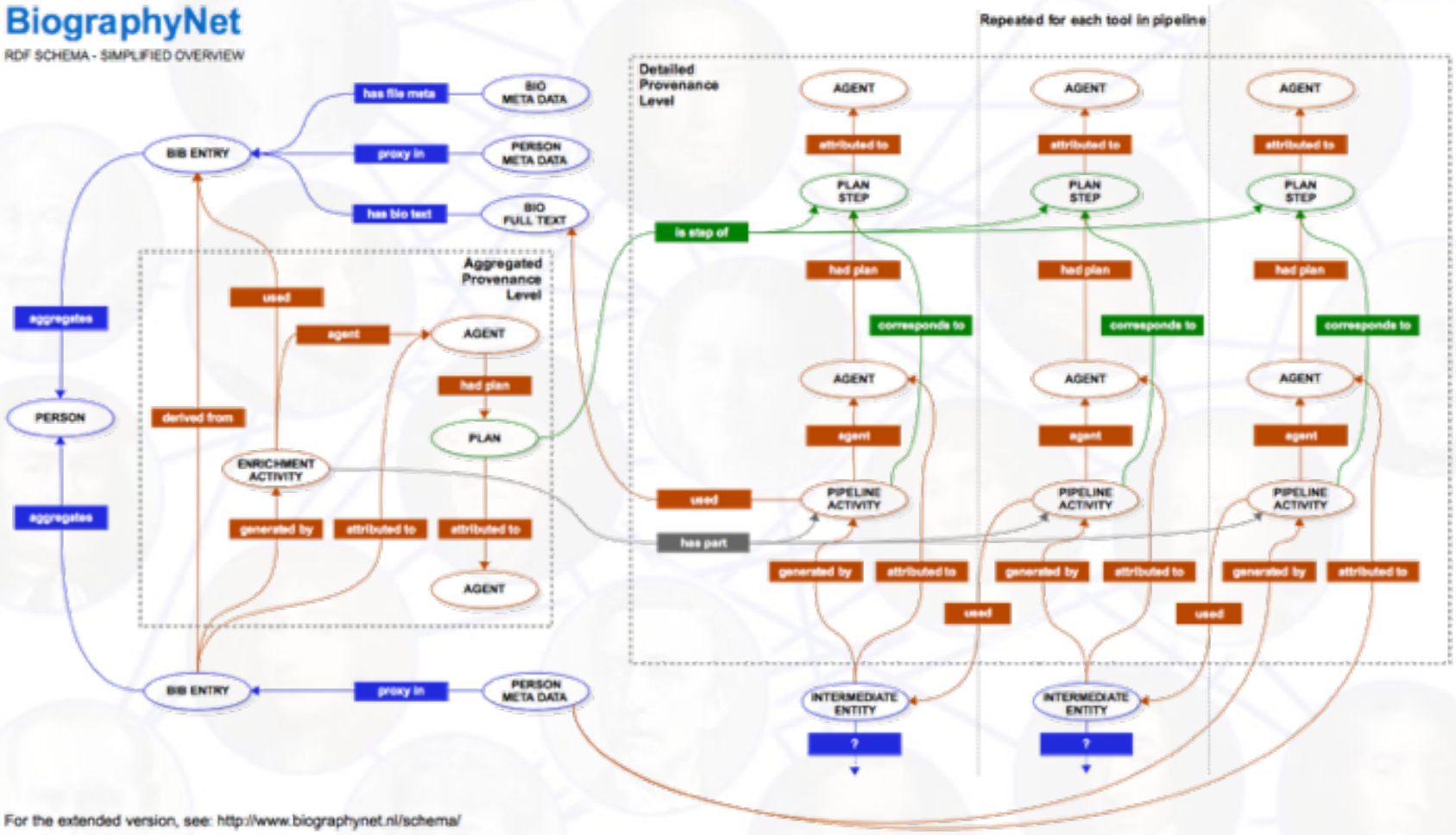
Provenance in BiographyNet

Needed to ensure ***credibility*** of the demonstrator, to ***evaluate*** its performance and to improve the ***academic status*** of the tool

- From several perspectives:
  - Information involved → Sources, but also: NER input data, etc.
  - Processes involved → All steps in enrichment, aggregation, etc
  - People involved → Who was responsible for pipeline, tool, etc.
- At multiple levels:
  - An ***aggregated level***, i.e. per enrichment → Targeted at the Historian
  - A ***detailed level***, i.e. all individual processes → Targeted at the Computer Scientist and computational linguist
- Including P-PLAN:* To not only model what ***actually happened***, but also what ***was supposed to happen***
  - Provides abstract information on idea behind activity, heuristics, assumptions, etc.
  - Allows for comparing the actual activity and its input/output with the original plan and its variables

*Daniel Garijo, Yolanda Gil; http://www.opmw.org/model/p-plan

RDF schema


BiographyNet
RDF SCHEMA - SIMPLIFIED OVERVIEW
Repeated for each tool in pipeline
Detailed Provenance Level
Aggregated Provenance Level
BIB ENTRY
has file meta
BIO META DATA
proxy in
PERSON META DATA
has bio text
BIO FULL TEXT
aggregates
PERSON
aggregates
BIB ENTRY
proxy in
PERSON META DATA
used
agent
AGENT
had plan
PLAN
derived from
ENRICHMENT ACTIVITY
generated by
attributed to
attributed to
AGENT
is step of
AGENT
attributed to
PLAN STEP
had plan
corresponds to
AGENT
agent
PIPELINE ACTIVITY
used
has part
generated by
attributed to
used
INTERMEDIATE ENTITY
?
For the extended version, see: http://www.biographynet.nl/schema/

Evaluation

- Two fold:
  - Building blocks
  - Historians questions

Building blocks

- Text annotations:
  - entities
  - events
  - time expressions
  - target concepts
  - relations with target concepts
- Comparison to metadata:
  - Birth and death date
  - Gender

Historian's questions

- Occurrences of concepts & people
- Group analyses:
  - educational background
  - age when obtaining function
- Overall corpus statistics:
  - Men versus women
  - Horoscope of people
  - Focus on specific century

Lessons learned: what worked well

- Have people from various disciplines share an office
- Constantly share information
  - about what humanities scholars want and
  - what computer scientists can deliver
- Always keep the intrinsic/extrinsic evaluation in mind:
  - the most reliable outcome depends on the use case

Lessons learned: what worked well

- Design your model carefully:
  - Make sure historians can access the information they want
  - Make it as compatible as possible with existing data representations
  - Provide information about the reliability of the data where you can:
    - Provenance
    - Confidence scores of tools

Lessons learned: what we would do differently

- Start developing evaluation material **from day 1**
- Get a full basic system as soon as possible

if you have a basic system and the means to evaluate, you know exactly what you should invest in

For future projects?

- Methodological insights:
  - Reliability, evaluation methods, provenance modelling
- The 2-step approach:
  1) From text to linguistic analyses
  2) From linguistic analyses to SEM

For future projects?

- The BiographyNet schema, SEM and GAF:
  - Event centric representation that is highly flexible
  - The schema explicitly captures provenance information
  - The schema is compatible with the Europeana data model

For future projects?

- The NLP tools:
  - Similar pipelines for linguistic analyses exist for English, Italian and Spanish
  - The interpretation software is only partially language specific
- D2D and the demonstrator are language independent:
  - D2D can handle anything represented in RDF
  - The demonstrator will be able to handle anything that uses the BN schema, SEM and GAF

The Future

- European project: extending to various data bases in different languages
- Common data structures for Biographical Data:
- Workshop on Digital Humanities 2016?

Rank	Individuals without their own biography	Number of mentions
1	Jezus Christus	> 75
2	Karel II (king of England)	60
3	Lodewijk XIV (king of France)	40
4	Lodewijk VIII (king of France)	25
5	Lodewijk XI (king of France)	25
6	Frans I (king of France)	23
7	Lodewijk XII (king of France)	18
8	Karl Marx (German philosopher)	18
9	Lodewijk XVI ((king of France)	16
10	Jozef II (German emperor)	15
11	Lodewijk XIII (king of France)	15
12	Napoleon Bonaparte (French emperor)	15

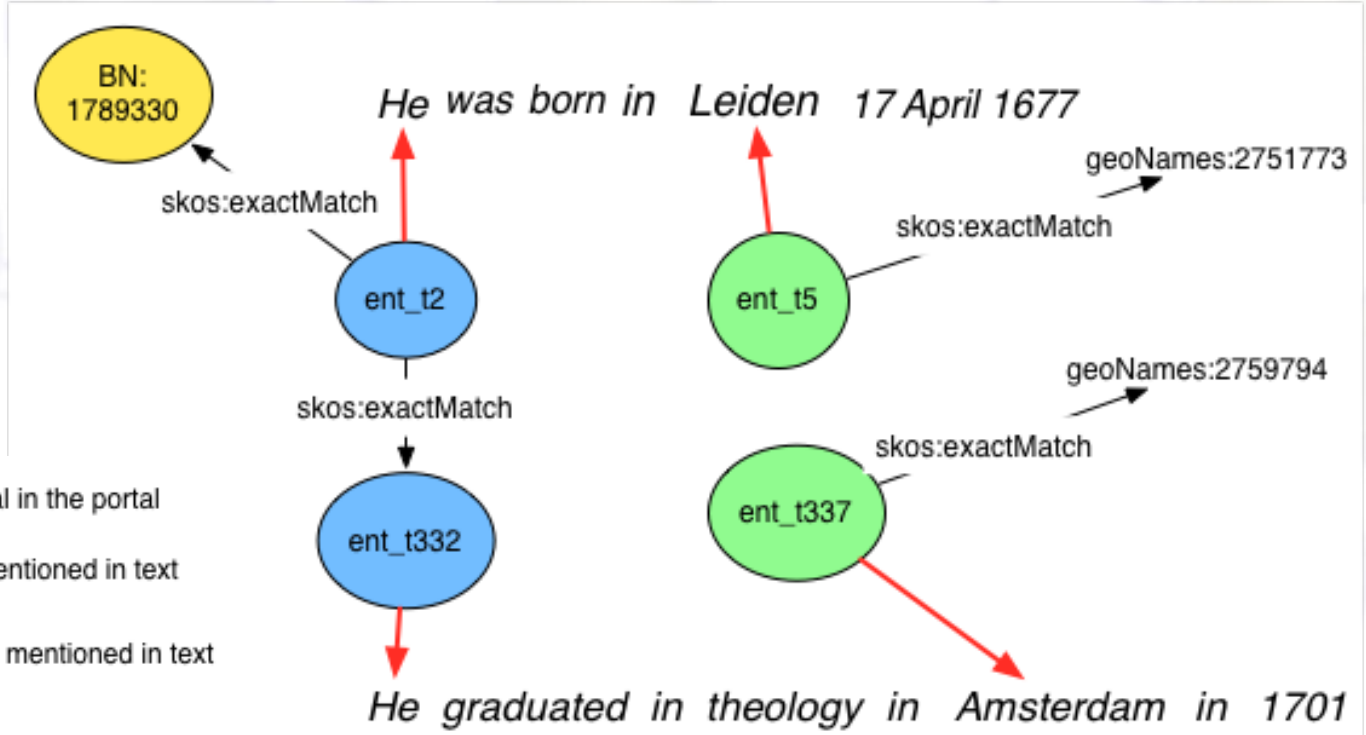
Table 5: People mentioned most frequently in the Biography Portal of the Netherlands, without their own biographical entry

Thank you!

Please visit: http://www.biographynet.nl

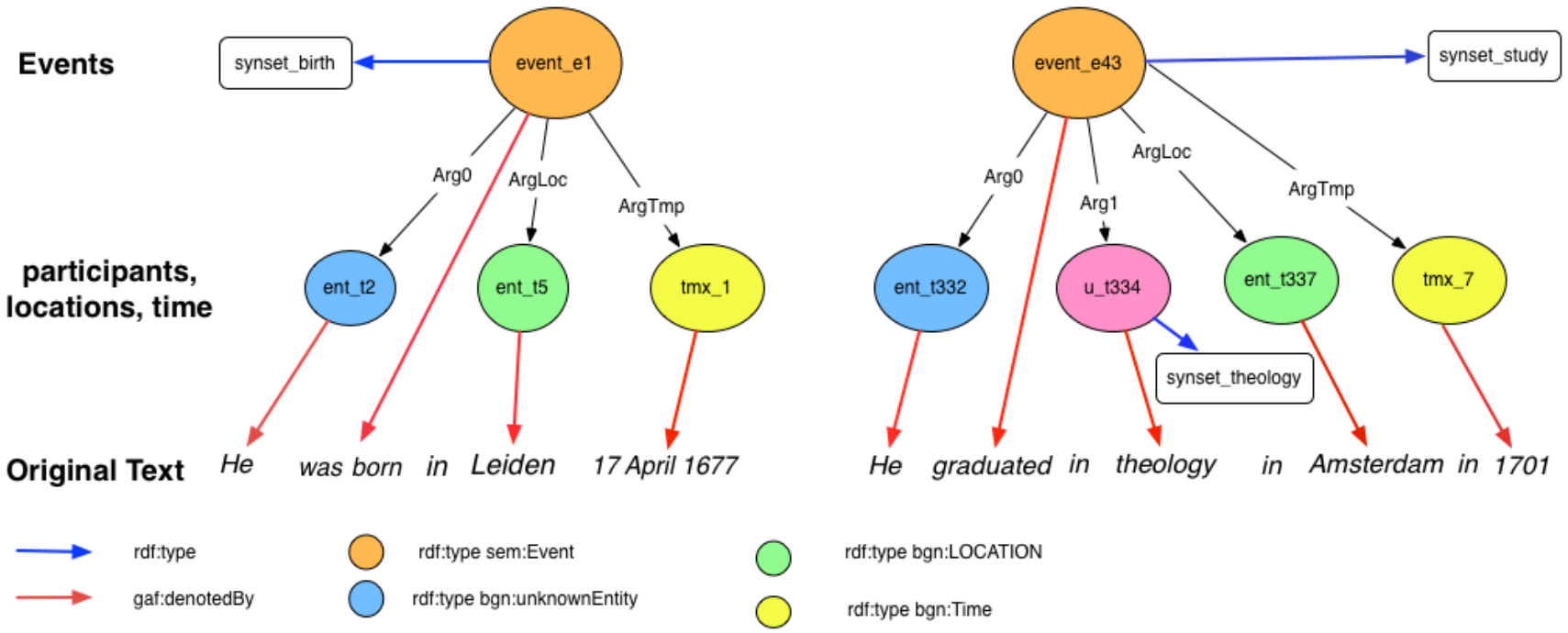
And let us know if we can help with anything!

Text Interpretation: Step 2



1. Which mentions refer to the same entity (and what is that)?
2. What is the exact geographic location that is meant?

Text Interpretation: step 1



Text interpretation (example)

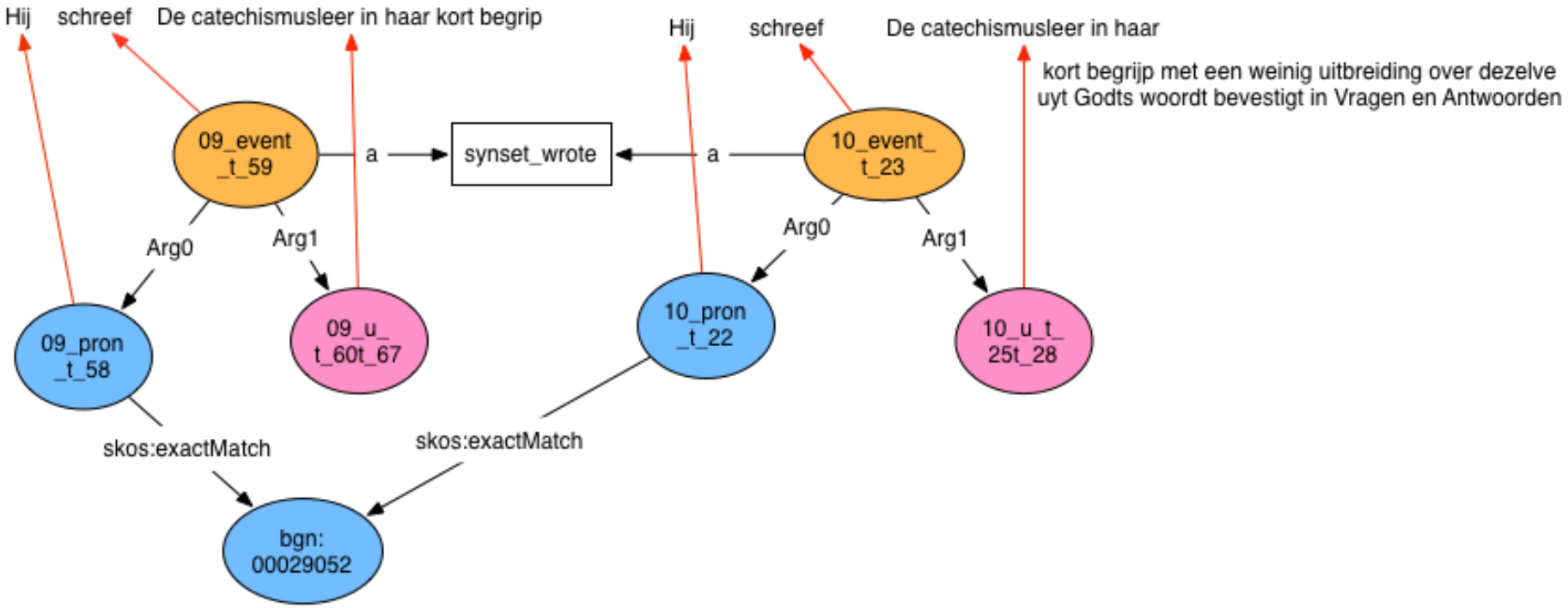
TIJKEN (Jacobus), geb. te Rotterdam 6 Febr. 1706, overl. te Amsterdam 1789. **Hij studeerde te Leiden** (ingeschr. 30 Juni 1723), werd 6 Febr. predt. te IJselmonde, 19 Juni 1735 te Schiedam, 8 Jan. 1741 te Amsterdam. Hij/was gehuwd met Agatha de Waerdt. **Hij schreef De catechismusleer in haar kort begrip, Amst.** 8o. Eene voorrede vr N. Barenzonius, Ziele des evangeliums 17442. Eene lykpredikatie op prins Willem IV. Schilderij door A. Folkema. Prenten door J. Folkema, J. Houbraken. Zie: Croese, Kerkel. Reg. 245. L. Knappert

NNBW

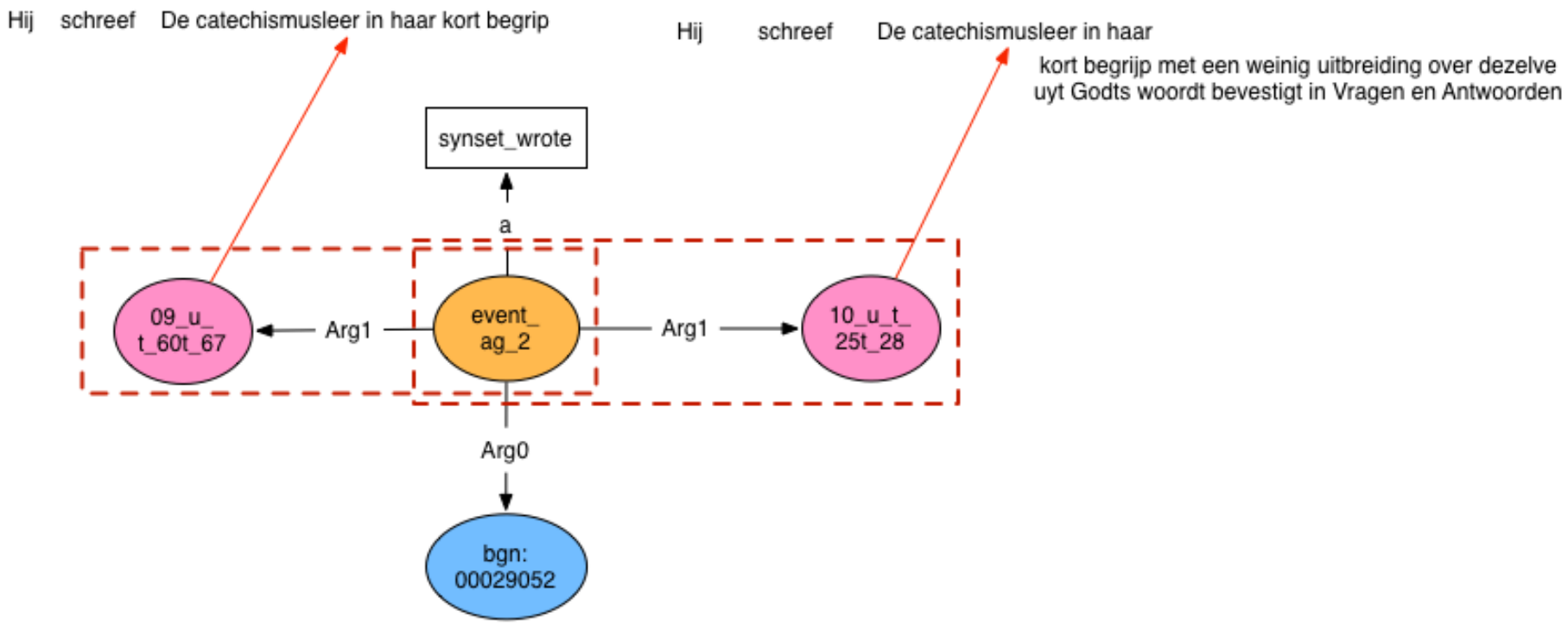
TIJKEN (Jacobus), predikant te IJsselmonde (1728), Schiedam (1735) en Amsterdam (1741). **Hij schreef: De Catechismusleer in haar kort begrijp met een weinig uytbreiding over deselve uyt Godts woordt bevestigt in Vragen en Antwoorden. 8o. Amst.**, 1750. 4de dr. Zie Koecher, Hist. d. Heidelb. catech., bl. 329, 330; Pauw en Veeris, Vern. kerk. Alphab., bl. 288; Abcoude, Tweede Aanh., bl. 151.

VDAA

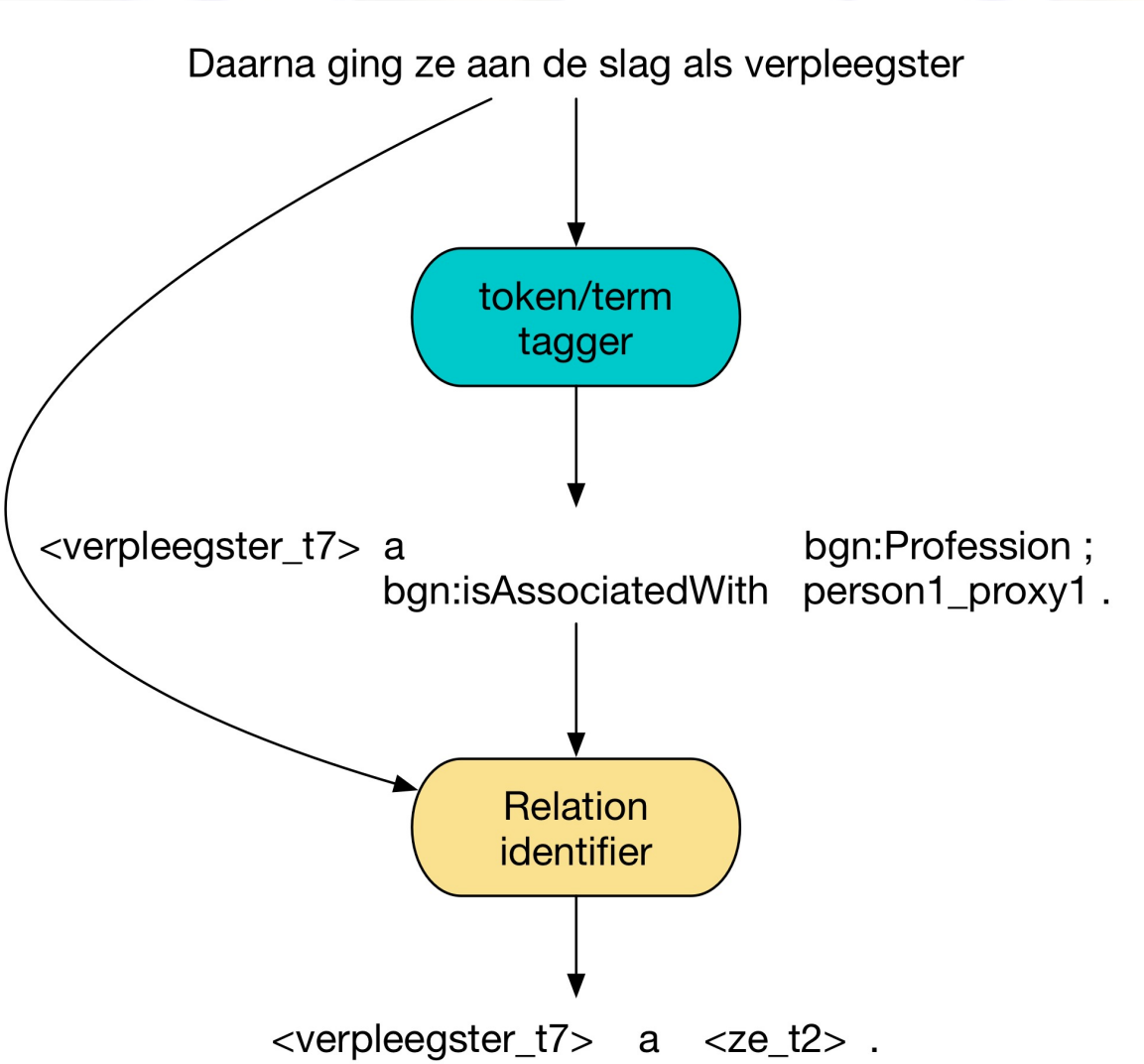
Text interpretation (example)



Text interpretation (example)



Interpretation: beyond events



Methods and Challenges

- Domain adaptation:
  - Domain specific meaning: *promoveren*
    typical biographical meaning: `getting a PhD'
    identified meaning includes: `change position on a scale'
  - HeidelTime: developed for the *Biographical dictionary of Socialism and Workers (BWSA)*
    - BWSA (late 20th century): 90.4% recall, 98.1% precision
    - BWG (late 20th century): 83% recall, 76,5% precision
    - VDAA (late 19th century): 69.7% recall, 77.6% precision

Approaches

- Targeted identification:
  - Concepts and events related to career
  - Pattern identification:
    - Dictionary specific patterns
    - Common structures

- Tool adaptation (most relevant)
  - Corpus specific abbreviations
  - Temporal expression variations

D2d Evaluation result details

	experiment:	E1	E2								
	statement	avg.	←agree 1	2	3	4	disagree→ 5	avg.	strong or weak agree (≤ 2)	p-value	not disagree (≤ 3)	p-value
S1	"Data 2 Documents seems to be a suitable approach to perform general Web Content Management such as the creation, sharing and placing of content articles"	2.14	17	40	9	6	1	2.10	57	$* < 10^{-6}$	66	$* < 10^{-12}$
S2	"Data 2 Documents seems to be a suitable approach to eliminate the traditional boundaries for Content Management between separate web sites, documents, and domains"	1.57	24	29	13	5	2	2.07	53	$* < 10^{-4}$	66	$* < 10^{-12}$
S3	"Data 2 Documents makes it easy to share content between separate web sites/documents/domains"	1.43	28	22	16	5	2	2.05	50	* 0.0011	66	$* < 10^{-12}$
S4	"Data 2 Documents seems to be a suitable approach to use Linked Data in web documents"	1.29	29	29	7	8	0	1.92	58	$* < 10^{-6}$	65	$* < 10^{-11}$
S5	"Manually editing Data 2 Documents definitions is not significantly harder to do than manually editing HTML"	2.29	25	15	18	13	2	2.34	40	0.2414	58	$* < 10^{-6}$
S6	"I would consider using Data 2 Documents, if I have to develop a general website in the future"	3.29	11	22	24	11	5	2.68	33	0.8254	57	$* < 10^{-6}$
S7	"I would consider using Data 2 Documents, if I have to develop a website in the future that makes use of Linked Data"	1.71	25	28	9	9	2	2.11	53	$* < 10^{-4}$	62	$* < 10^{-9}$

This table shows an aggregation of the answers of the participants of the two experiments (E1 and E2) about the degree to which they agree with the given statements about the usability and usefulness of d2d. The range was from strongly agree (1) to strongly disagree (5). Statistically significant p-values are marked with *.

netherlands eScience center

huygens ing

VU UNIVERSITY AMSTERDAM