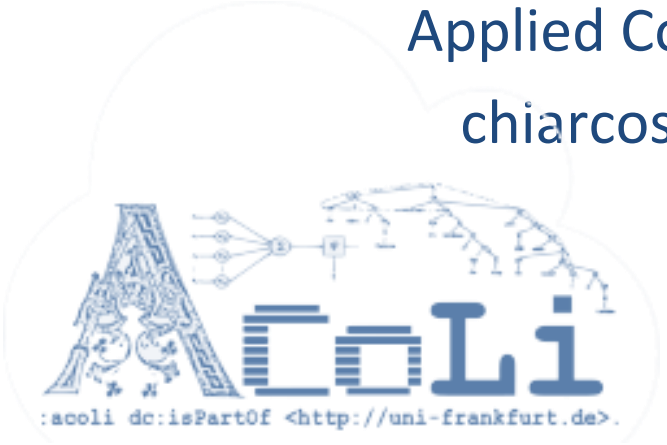
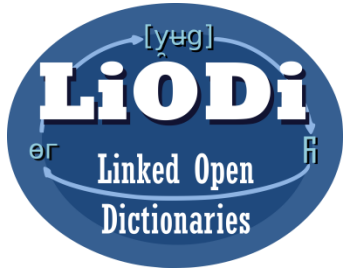
Web Annotation (W3C Standard)

Christian Chiarcos

Applied Computational Linguistics (ACoLi)

chiarcos@informatik.uni-frankfurt.de



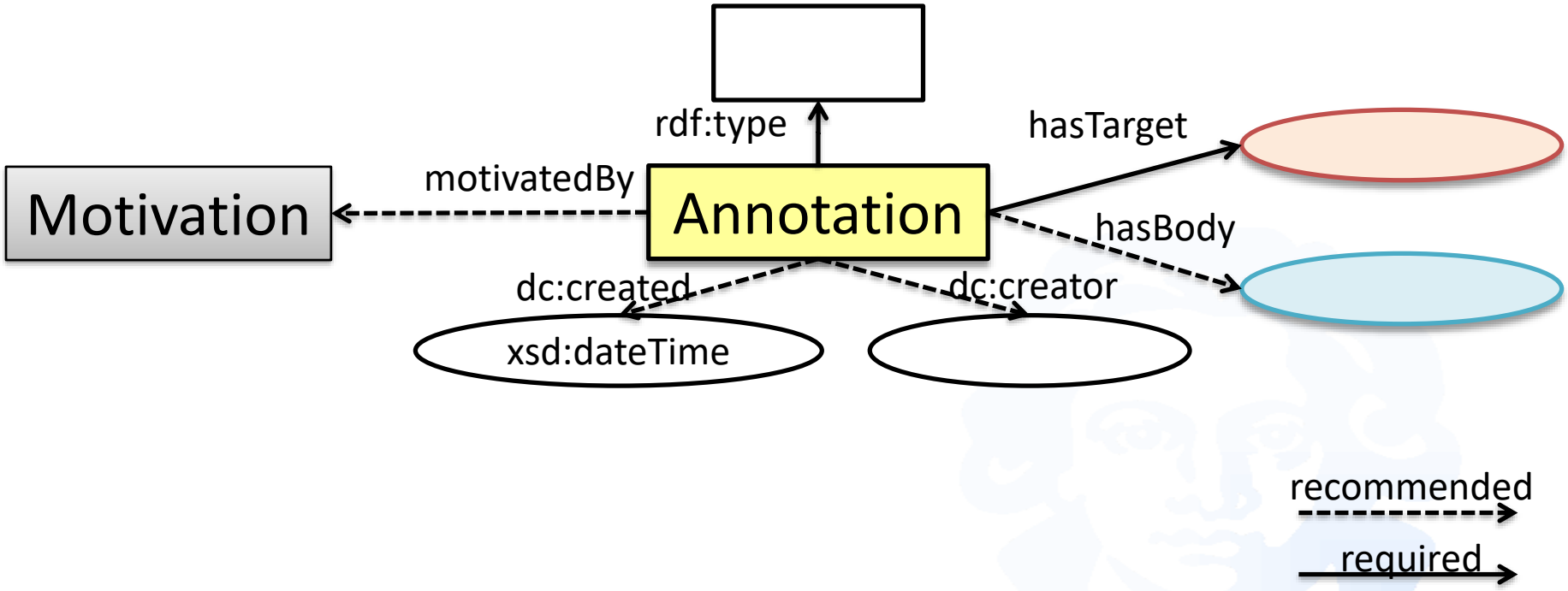


Web Annotation / Open Annotation

- W3C Open Annotation Community Group
  - https://www.w3.org/community/openannotation/
  - 2012-2014
    - mostly driven by bioinformatics, but generic formalism for annotating web content
- Web Annotation (W3C recommendations, Feb 2017)
  - Data Model: https://www.w3.org/TR/annotation-model
    - general description
  - Vocabulary: https://www.w3.org/TR/annotation-vocab
    - ontology
  - Protocol: https://www.w3.org/TR/annotation-protocol
    - retrieving and manipulating annotations
    - serialization: *must* JSON-LD, *should* Turtle, *may* provide other RDF serializations

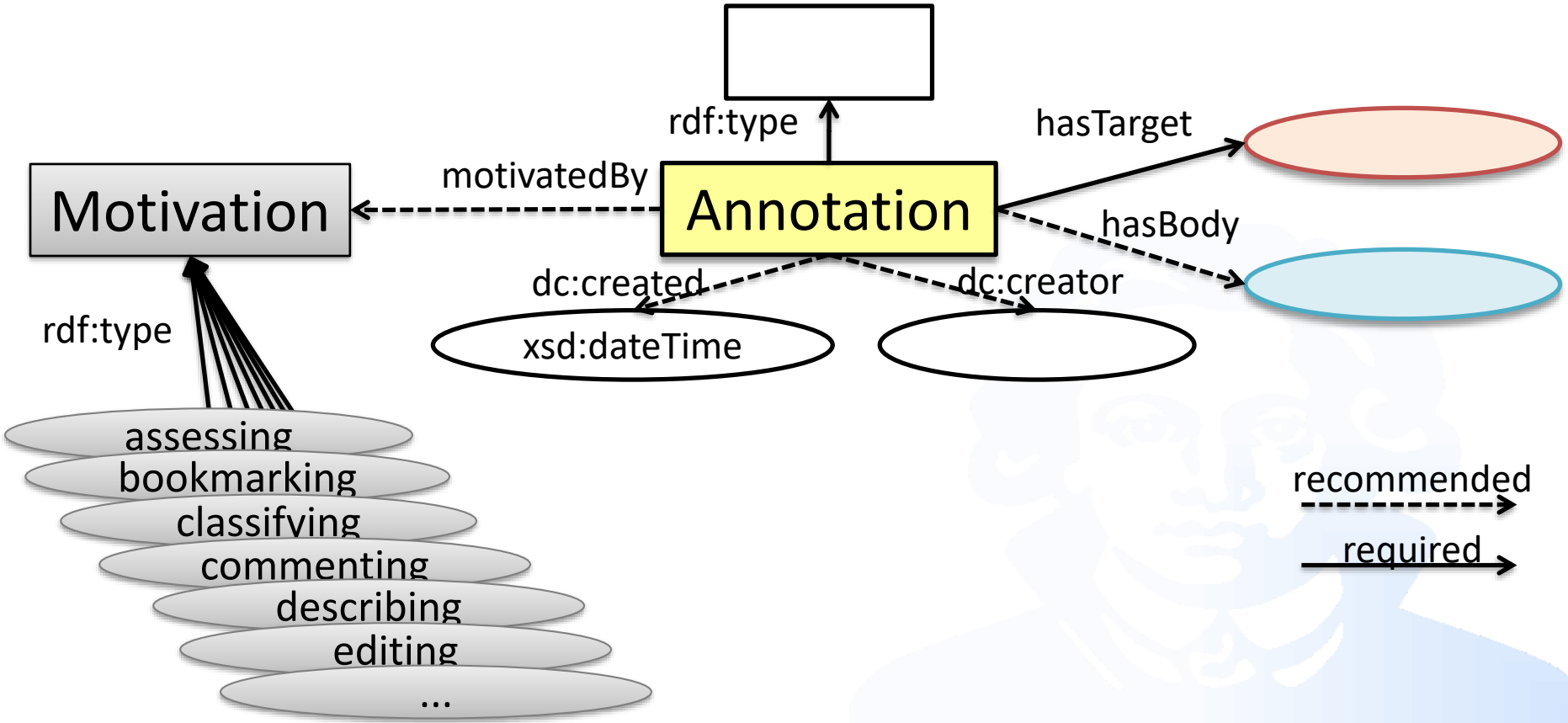
Web Annotation: Annotation

https://www.w3.org/TR/annotation-vocab/



Web Annotation: Annotation

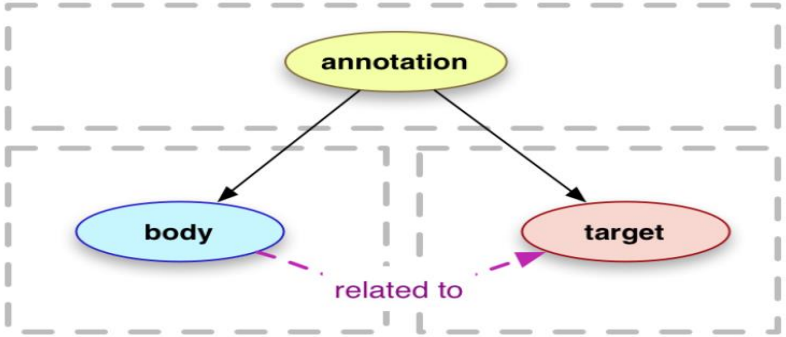
https://www.w3.org/TR/annotation-vocab/



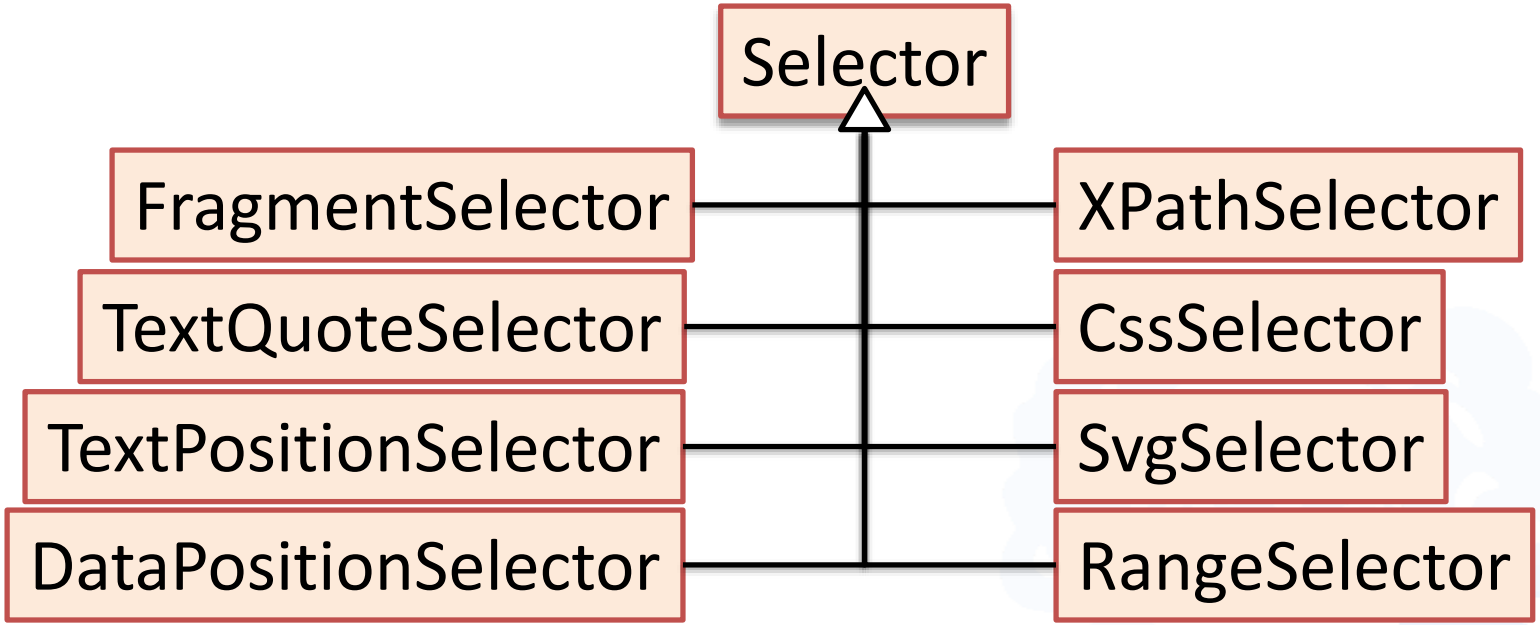
Web Annotation: Target and Body

https://www.w3.org/TR/annotation-model/

- body
  - element containing the annotation
  - object property: *oa:hasBody* (any RDF object)
  - datatype property: *oa:bodyValue* (strings)
- target
  - element being annotated
  - any RDF object, *including*
    - oa:Selector (more in a second)



oa:Selector – *e.g. possible targets*



or: just any URI ;)

e.g., a String URI (RFC5147)

http://example.com/text.txt#char=100,105

Named Entity Annotations (ENAMEX)

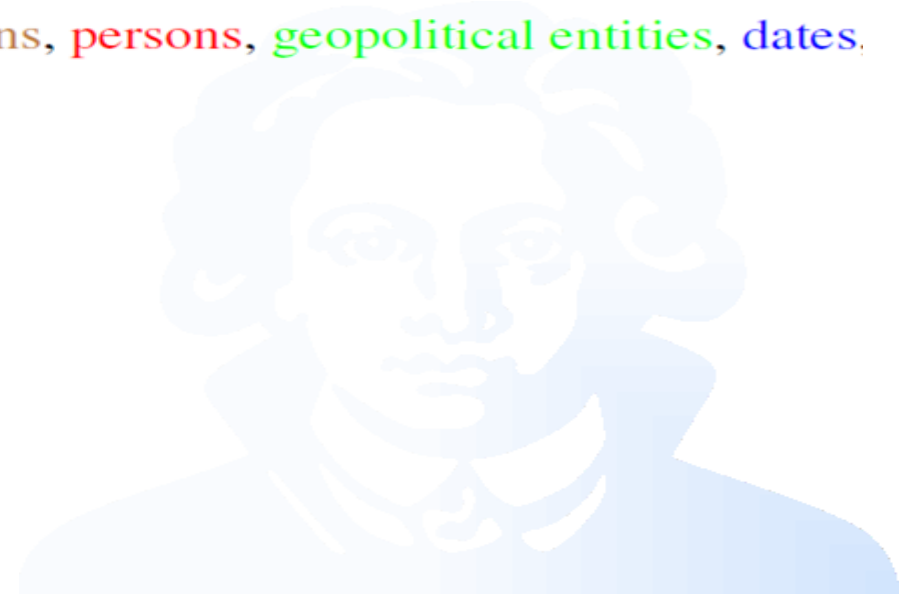
Secretary of State James Baker, who accompanied President Bush to Costa Rica, told reporters Friday: "I have no reason to deny reports that some Contras ambushed some Sandinista soldiers. "

OntoNotes corpus, wsj-0655

https://catalog.ldc.upenn.edu/LDC2013T19

James	B-PERSON
Baker	E-PERSON
told	O
reporters	O
Friday	S-DATE
:	O

organizations, persons, geopolitical entities, dates.



Named Entity Annotations (JSON-LD)

Secretary of State James Baker, who accompanied President Bush to Costa Rica, told reporters Friday: "I have no reason to deny reports that some Contras ambushed some Sandinista soldiers. "

```
James       B-PERSON
Baker       E-PERSON
told        O
reporters   O
Friday      S-DATE
:           O
```

```
{
  "@graph": [
        {
          "@context": "http://www.w3.org/ns/anno.jsonld",
          "id": "http://example.org/enamex2",
          "type": [
             "Annotation",
                 "https://catalog.ldc.upenn.edu/docs/LDC2007T21/
                     ontonotes-1.0-documentation.pdf#ENAMEX"
          ],
          "body": {
                "type" : "TextualBody",
                "value" : "PERSON",
                "format" : "text/plain"
          },
          "target": {
                "source": "https://catalog.ldc.upenn.edu/
                    ldc2013t19/data/files/data/english/
                    annotations/nw/wsj/06/wsj_0655.name",
                "selector": {
                  "type": "TextQuoteSelector",
                  "exact": "James Baker"
                }  }
        }
] }
```

Named Entity Annotations (Turtle)

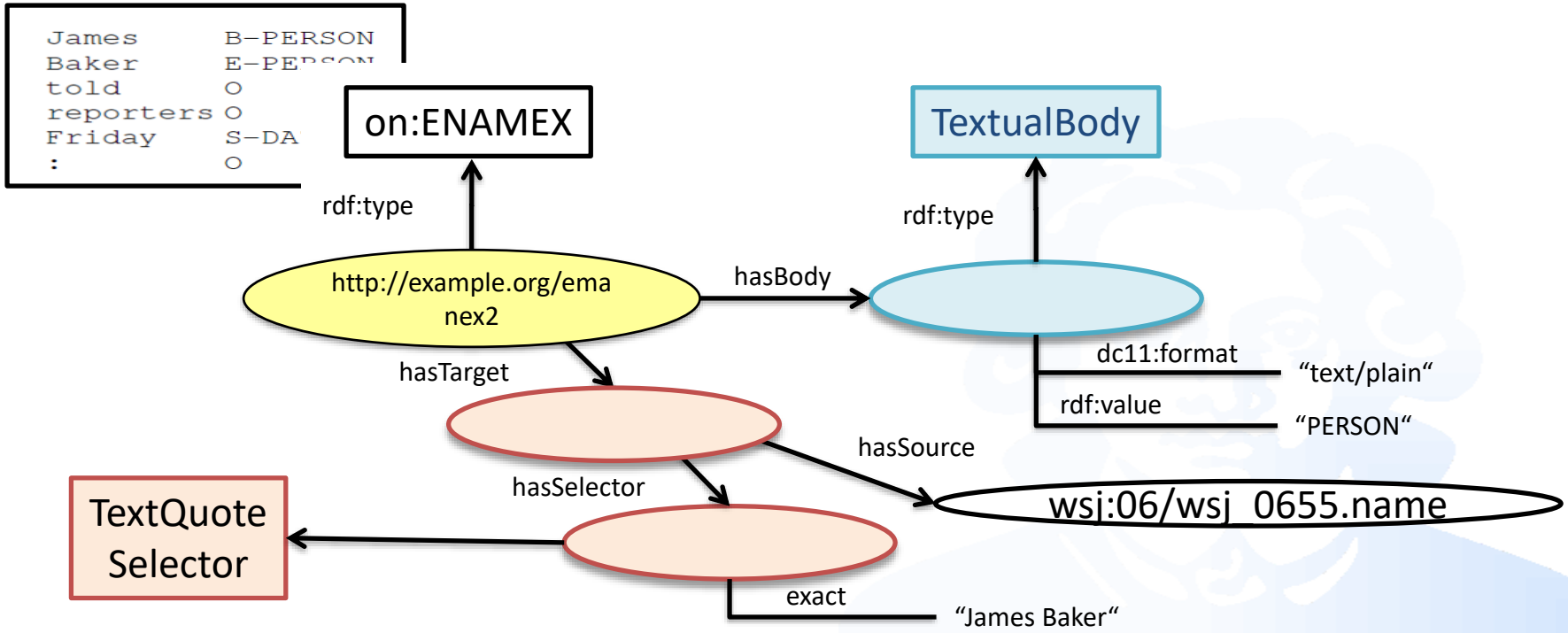
Secretary of State James Baker, who accompanied President Bush to Costa Rica, told reporters Friday: "I have no reason to deny reports that some Contras ambushed some Sandinista soldiers. "

James	B-PERSON
Baker	E-PERSON
told	O
reporters	O
Friday	S-DATE
:	O

```
<http://example.org/enamex2>
  a oa:Annotation, on:ENAMEX ;
  oa:hasBody [
    a oa:TextualBody ;
    dc11:format "text/plain"^^xsd:string ;
    rdf:value "PERSON"^^xsd:string
  ] ;
  oa:hasTarget [
    oa:hasSelector [
      a oa:TextQuoteSelector ;
      oa:exact "James Baker"^^xsd:string
    ] ;
    oa:hasSource wsj:06/wsj_0655.name
  ] .
```

Named Entity Annotations

Secretary of State James Baker, who accompanied President Bush to Costa Rica, told reporters Friday: "I have no reason to deny reports that some Contras ambushed some Sandinista soldiers. "



Web Annotation: Overview

- relatively good uptake
  - esp. in bioinformatics
- reification
  - annotation as *n:m* relation between bodies & targets
  - with metadata
- powerful
  - annotate all instances of a string at once using a *oa:TextQuoteSelector*
- very verbose
  - „X is a person according to *on:ENAMEX* annotations“ takes 11 triples

Named Entity Annotations (Turtle)

Secretary of State James Baker, who accompanied President Bush to Costa Rica, told reporters Friday: "I have no reason to deny reports that some Contras ambushed some Sandinista soldiers. "

James	B-PERSON
Baker	E-PERSON
told	O
reporters	O
Friday	S-DATE
:	O

```
<http://example.org/enamex2>
  a oa:Annotation, on:ENAMEX ;
  oa:hasBody [
    a oa:TextualBody ;
    dc11:format "text/plain"^^xsd:string ;
    rdf:value "PERSON"^^xsd:string
  ] ;
  oa:hasTarget [
    oa:hasSelector [
      a oa:TextQuoteSelector ;
      oa:exact "James Baker"^^xsd:string
    ] ;
    oa:hasSource wsj:06/wsj_0655.name
  ] .
```

notational shorthand: *oa:bodyValue* for string-value bodies

3-7 oa:bodyValue "PERSON"^^xsd:string ;

=> 4 triples replaced by 1

Web Annotation: Overview

- relatively good uptake
  - esp. in bioinformatics
- reification
  - annotation as *n:m* relation between bodies & targets
  - with metadata
- powerful
  - annotate all instances of a string at once using a *oa:TextQuoteSelector*
- rather verbose
  - „X is a person according to *on:ENAMEX* annotations“ takes ~~11~~ 7 triples

Web Annotation: Overview

- relatively good uptake
  - esp. in bioinformatics
- reification
  - annotation as *n:m* relation between bodies & targets
  - with metadata
- powerful
  - annotate all instances of a string at once using a *oa:TextQuoteSelector*
- rather verbose
- no linguistic data structures

Tooling: Recogito

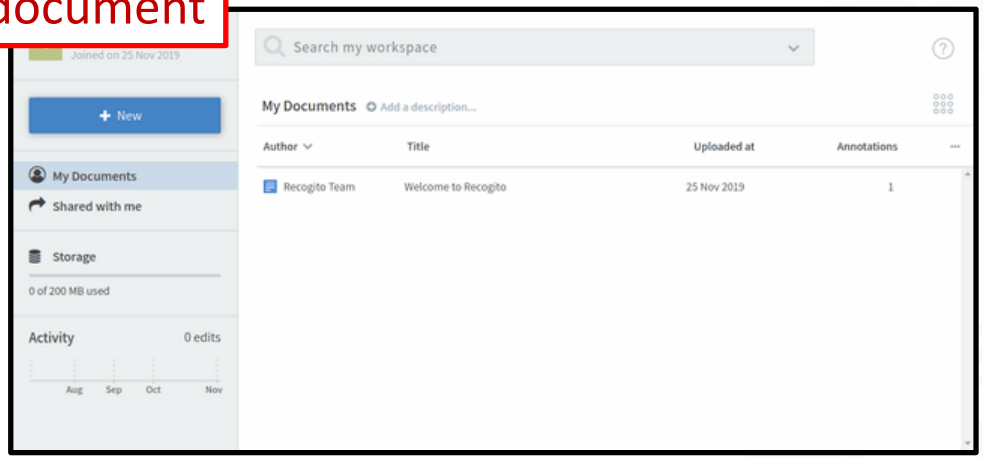
https://recogito.pelagios.org

- Recogito is a tool designed for the annotation of maps and texts with geographical entities
- Developed by the Pelagios network
  - a long-running initiative that links information online through common references to places
  - Digital Humanities, e.g., historical maps, excavation reports, literature
- Based on Web Annotation
  - can be applied to numerous source formats
  - limited to labelling and linking tasks

Recogito

- fully graphical
- underlying technology hidden from the user

1. upload a document



all screenshots from https://recogito.pelagios.org/help/tutorial

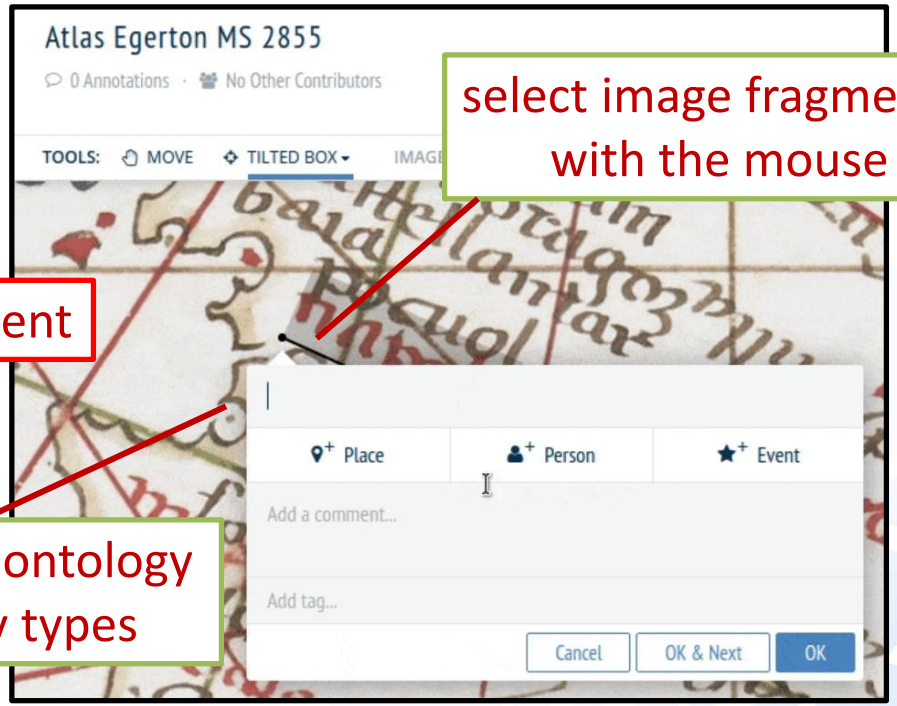
Recogito

- fully graphical

1. upload a document

2. annotate your document

e.g., image



select image fragments with the mouse

underlying ontology of entity types

all screenshots from https://recogito.pelagios.org/help/tutorial

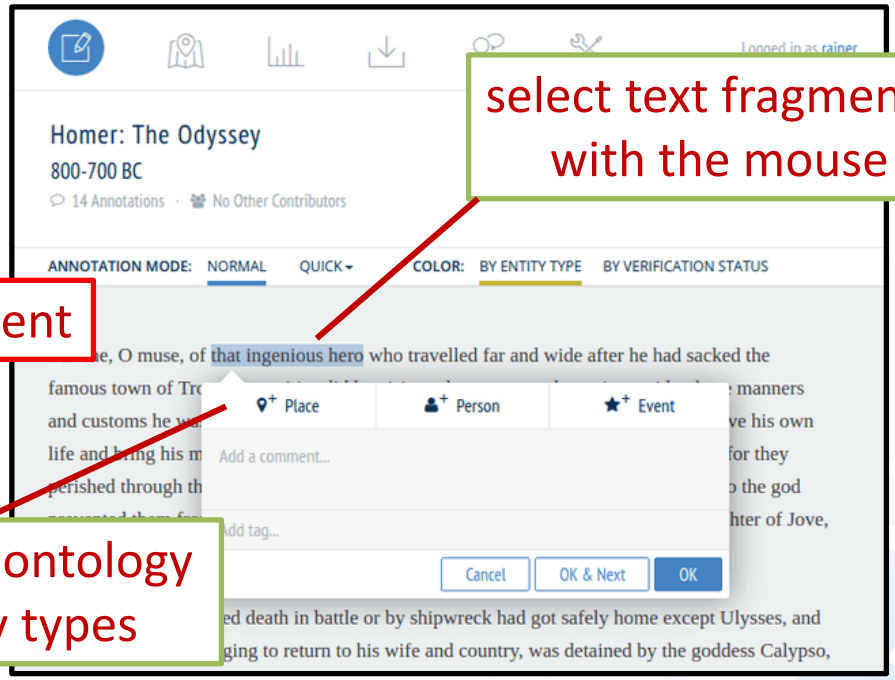
Recogito

- fully graphical

1. upload a document

2. annotate your document

e.g., image or text



select text fragments with the mouse

underlying ontology of entity types

all screenshots from https://recogito.pelagios.org/help/tutorial

Recogito

- fully graphical

1. upload a document
2. annotate your document
3. link with Gazeteers

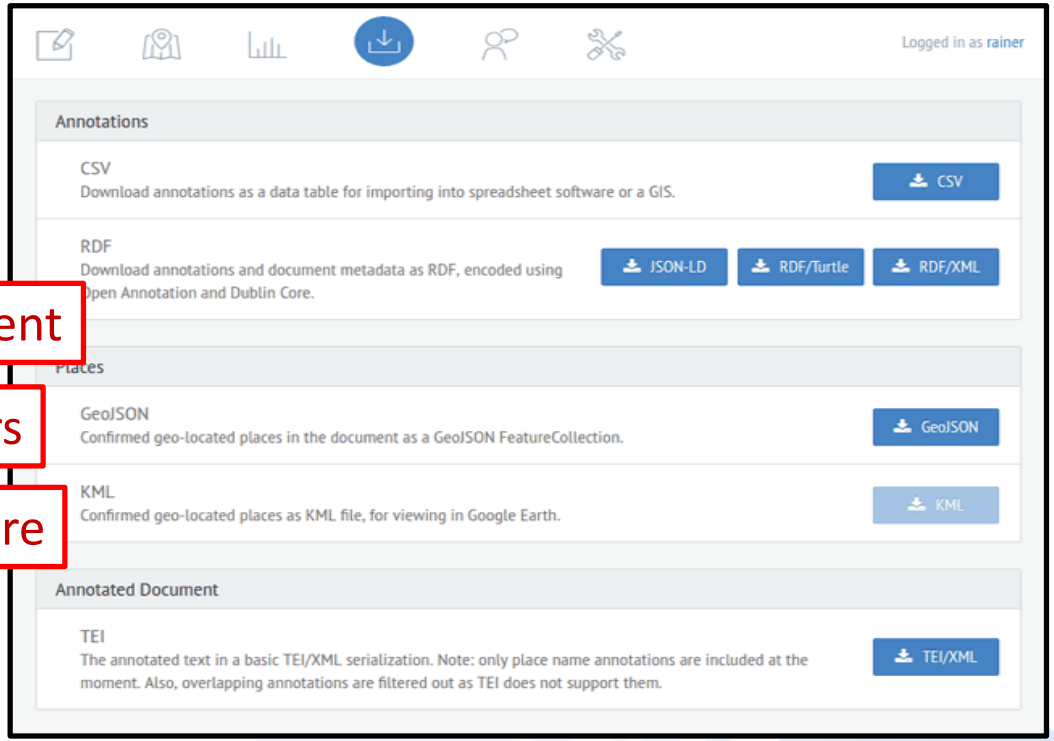
annotation integrates special-purpose views for Geoinformation

all screenshots from https://recogito.pelagios.org/help/tutorial

Recogito

- fully graphical

1. upload a document
2. annotate your document
3. link with Gazeteers
4. export and share



illustrates the potential to develop applications for end users

... the tool itself is for a very restricted use case

Web Annotation

Strengths

- very generic
- multi-modal
- extensible
- full-fledged W3C standard
- relatively widely used
  - esp., bioinformatics, DH

Weaknesses

- labelling & linking, only
- no linguistic data structures
  - phrases?
  - relations?
- very verbose
- counter-intuitive terminology