Metadata Enrichment Tools

Sep 16, 2022 | Weave Toolbox Presentation

Orfeas Menis

Co-financed by the European Union
Connecting Europe Facility

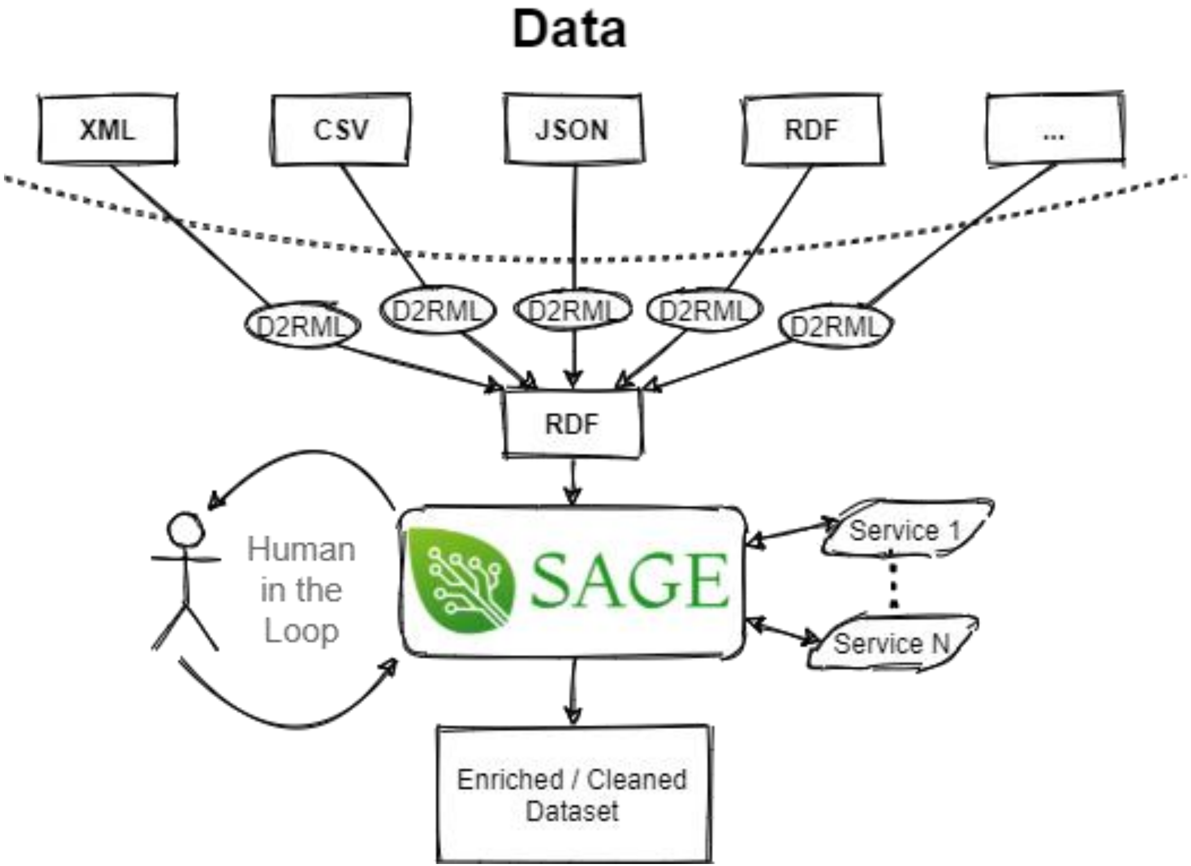
The Platform

SAGE is a Semantic Enrichment and Data management platform

It can import Heterogeneous Data (e.g. XML, CSV, JSON, RDF, etc.) from multiple sources.

Transform and Enrich the data through external services.

Overview and improvement through a Human in the Loop approach.



Metadata Enrichment

- Metadata enrichment should:
  - enable the creation of links between objects and contextual entities (persons, places, concepts and timespans)
  - allow the representation of (real-world) entities related to a provided object as fully fledged resources, not just strings
- Link to knowledge bases like Wikidata, Getty, Geonames, Europeana Thesauri, and others (Linked Data).

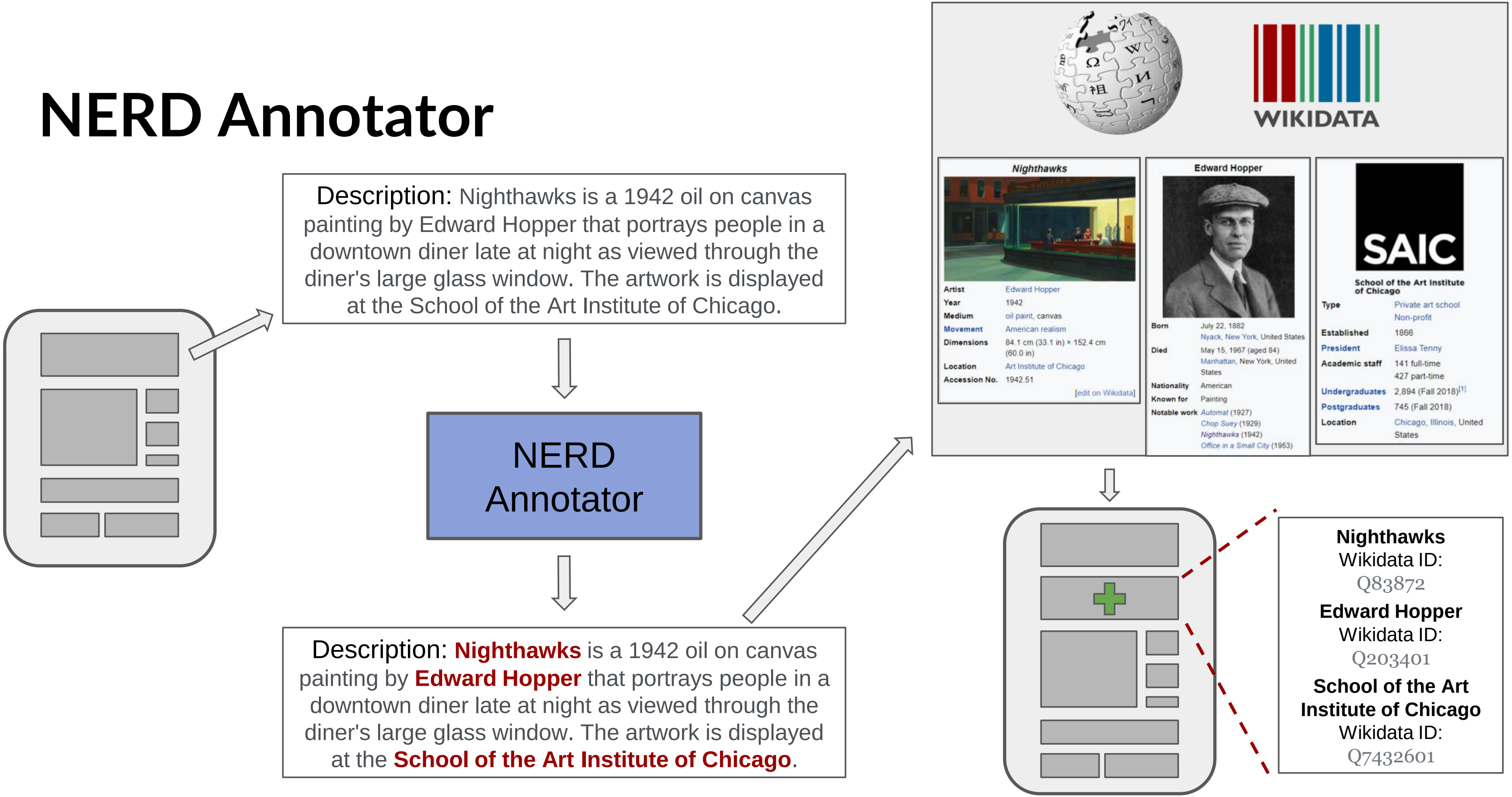
NERD Annotator

- Perform Named Entity Recognition and Disambiguation (NERD)
- Links to Wikidata
- Better performance on longer texts with rich context
- Plug-n-play solution, no fine-tuning needed

Technologies:

- Semantic-based Natural Language Processing Technologies employing Knowledge Graphs, etc.
- Named Entity Recognition and Disambiguation (NERD) techniques

NERD Annotator



Description: Nighthawks is a 1942 oil on canvas painting by Edward Hopper that portrays people in a downtown diner late at night as viewed through the diner's large glass window. The artwork is displayed at the School of the Art Institute of Chicago.

NERD Annotator

Description: **Nighthawks** is a 1942 oil on canvas painting by **Edward Hopper** that portrays people in a downtown diner late at night as viewed through the diner's large glass window. The artwork is displayed at the **School of the Art Institute of Chicago**.

Nighthawks

Artist	Edward Hopper
Year	1942
Medium	oil paint, canvas
Movement	American realism
Dimensions	84.1 cm (33.1 in) × 152.4 cm (60.0 in)
Location	Art Institute of Chicago
Accession No.	1942.51

[edit on Wikidata]

Edward Hopper

Born	July 22, 1882 Nyack, New York, United States
Died	May 15, 1967 (aged 84) Manhattan, New York, United States
Nationality	American
Known for	Painting
Notable work	*Automat* (1927) *Chop Suey* (1929) *Nighthawks* (1942) *Office in a Small City* (1953)

School of the Art Institute of Chicago

Type	Private art school Non-profit
Established	1866
President	Elissa Tenny
Academic staff	141 full-time 427 part-time
Undergraduates	2,894 (Fall 2018)[1]
Postgraduates	745 (Fall 2018)
Location	Chicago, Illinois, United States

Nighthawks
Wikidata ID:
Q83872

Edward Hopper
Wikidata ID:
Q203401

School of the Art Institute of Chicago
Wikidata ID:
Q7432601

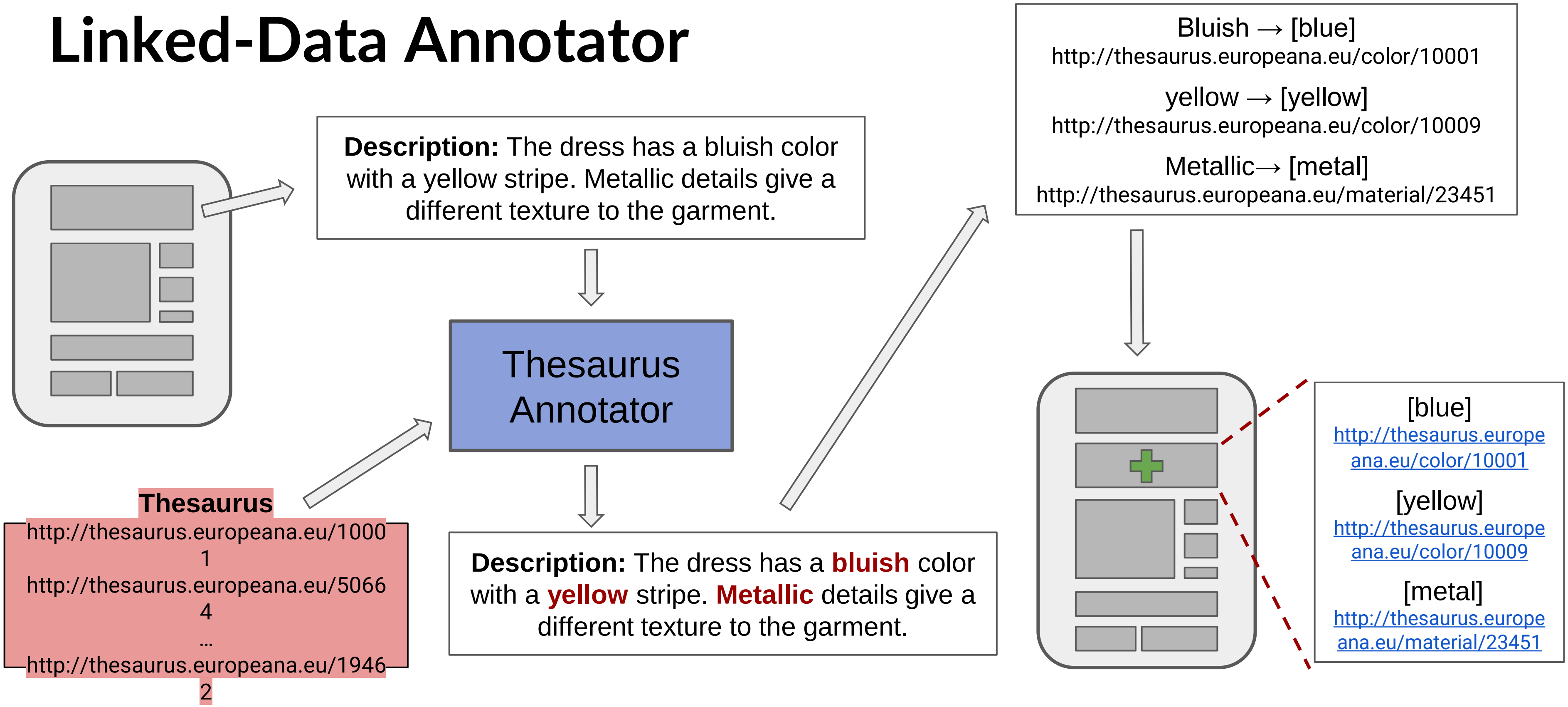
Linked-Data Annotator

- Link text to Thesaurus/Vocabulary terms
- Smart String Matchings utilizing state-of-the-art NLP technologies
- Improved time performance (time optimization)
- Use existing thesaurus/vocabulary or create custom from a list of keywords and the respective URIs

Technologies:

- Bert-based Natural Language Processing (NLP) models
- Lemmatizers
- RDF Thesauri

Linked-Data Annotator



Description: The dress has a bluish color with a yellow stripe. Metallic details give a different texture to the garment.

Thesaurus

http://thesaurus.europeana.eu/10001
http://thesaurus.europeana.eu/50664
…
http://thesaurus.europeana.eu/19462

Description: The dress has a **bluish** color with a **yellow** stripe. **Metallic** details give a different texture to the garment.

Bluish → [blue]
http://thesaurus.europeana.eu/color/10001

yellow → [yellow]
http://thesaurus.europeana.eu/color/10009

Metallic→ [metal]
http://thesaurus.europeana.eu/material/23451

[blue]
http://thesaurus.europeana.eu/color/10001

[yellow]
http://thesaurus.europeana.eu/color/10009

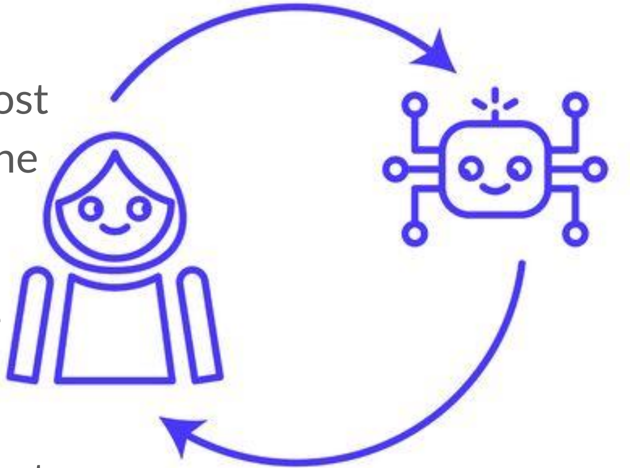
[metal]
http://thesaurus.europeana.eu/material/23451

Human in the Loop: Validate and Re-train

Humans validate the enrichments and **fine-tune the annotations** to meet the requirements.

Statistical Analysis from the Validation procedure providing most rejected annotations and searching for patterns to generalize the manual validation.

Improvement of the services from the **human feedback** (Active Learning). *[Experimental]*



Create exportation filters, filtering out any unwanted enrichments before exportation/publication based on the validation results.

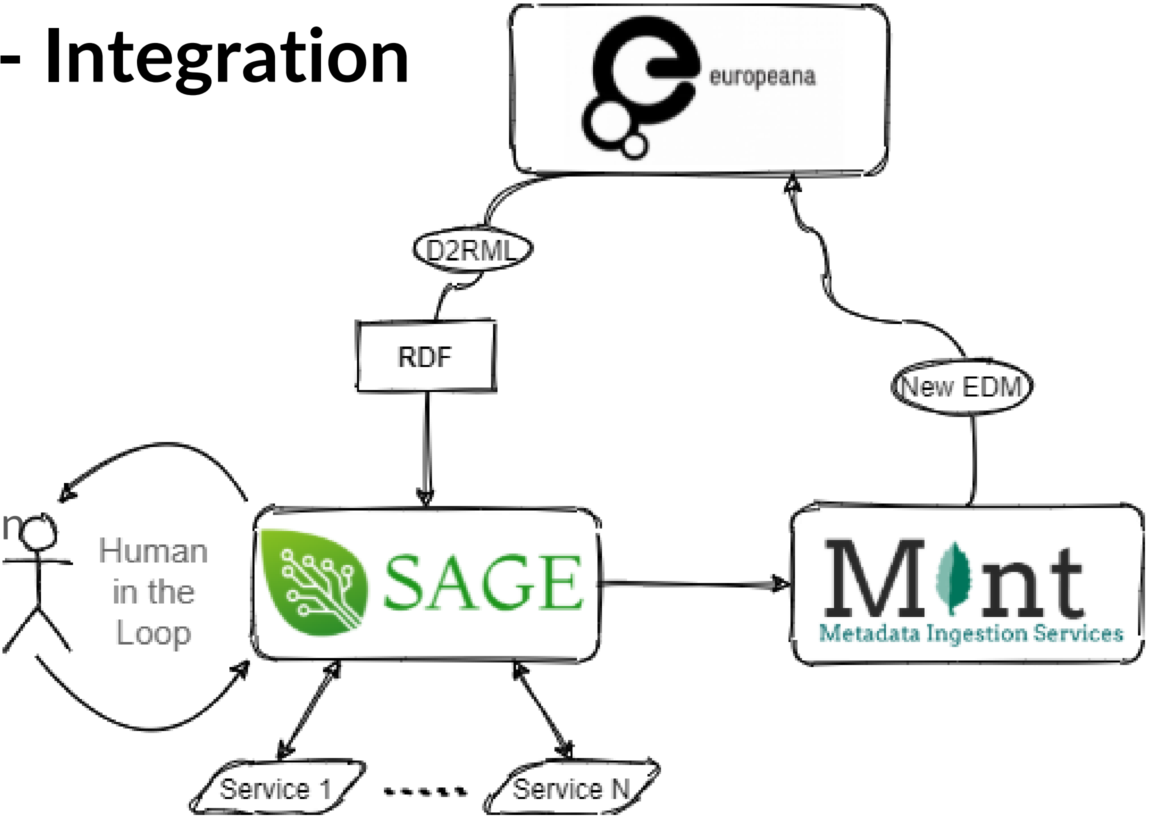
SAGE & Europeana - Integration

Harvest Data through the Europeana API

[New] Direct integration with MINT for data importation

New EDM to include an indication for machine generated metadata.

Publication back to Europeana through MINT.



SAGE & WEAVE in Numbers

Enrichments

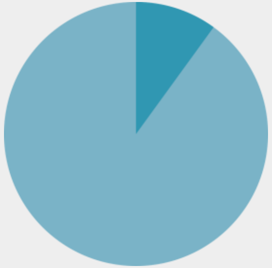
9621

Enriched Records

3849

Users Involved

10



Acceptance Rate

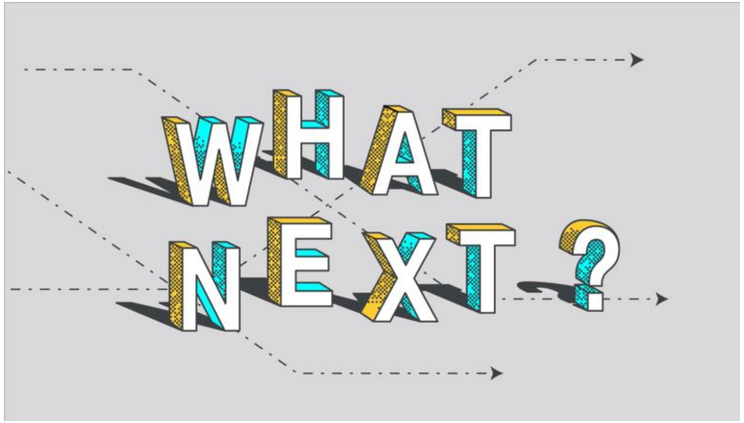
91.2%

Future Improvements / Directions

Create new custom services with state-of-the-art / cutting-edge NLP technologies.

Create a complete validation framework (open research area).

Redesign of complex parts of the platform in order to be easy to use even for non-technical persons.



Thank you!