Combining Automatic Annotation with Human Validation for the Semantic Enrichment of Cultural Heritage Metadata

Eirini Kaldeli, Alexandros Chortaras, Vassilis Lyberatos, Jason Liartis, Spyridon Kantarelis and Giorgos Stamou

AI and Learning Systems Lab, School of Electrical and Computer Engineering, National Technical University of Athens, Greece

Abstract

The addition of controlled terms from linked open datasets and vocabularies to metadata can increase the discoverability and accessibility of digital collections. However, the task of semantic enrichment requires a lot of effort and resources that cultural heritage organizations often lack. State-of-the-art AI technologies can be employed to analyse textual metadata and match it with external semantic resources. Depending on the data characteristics and the objective of the enrichment, different approaches may need to be combined to achieve high-quality results. What is more, human inspection and validation of the automatic annotations should be an integral part of the overall enrichment methodology. In the current paper, we present a methodology and supporting digital platform, which combines a suite of automatic annotation tools with human validation for the enrichment of cultural heritage metadata within the European data space for cultural heritage. The methodology and platform have been applied and evaluated on a set of datasets on crafts heritage, leading to the publication of more than 133K enriched records to the Europeana platform. A statistical analysis of the achieved results is performed, which allows us to draw some interesting insights as to the appropriateness of annotation approaches in different contexts. The process also led to the creation of an openly available annotated dataset, which can be useful for the in-domain adaptation of ML-based enrichment tools.

Keywords

semantic enrichment, cultural heritage metadata, named entity recognition and disambiguation

1. Introduction

Semantic enrichment is the process of adding new semantics to unstructured data, such as free text, so that machines can make sense of it and build connections to it. In the case of the metadata that describes Cultural Heritage (CH) items, unstructured data comes in the form of free text that details several aspects of the item, for example its main characteristics, its location, creator, etc. Through the process of semantic enrichment, those textual descriptions are analyzed and augmented with controlled terms from Linked Open datasets, such as Wikidata[1] and Geonames[2], or controlled vocabularies, such as the Getty Art & Architecture Thesaurus[3]

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

ekaldeli@image.ece.ntua.gr (E. Kaldeli)
0000-0001-7045-2588 (E. Kaldeli)

[1] https://www.wikidata.org/

[2] https://www.geonames.org/

[3] https://vocab.getty.edu/aat/

(AAT). Those terms represent concepts and attributes (e.g. “costume”, “Renaissance”, colors), named entities, such as persons, locations, and organisations, or chronological periods. For example, the strings “Leonardo da Vinci” and “da Vinci, Leonardo” can be both linked to the Wikidata term representing the Italian Renaissance polymath. This additional piece of information associated with a CH resource is commonly referred to as an *annotation*, which links the CH object with some URI (Unique Reference Identifier) derived from vocabularies or open data sources.

Semantic enrichment adds meaning and context to digital collections and makes them more easily discoverable. Given its importance, it has been a main concern and focus of efforts by the Europeana digital library[4] as well as individual data aggregators and providers. Firstly, linked data makes the meaning of textual metadata unambiguous [25]. For example, the string “Leonardo da Vinci” may refer, depending on the context, to the Italian Renaissance polymath or the homonymous airport in Fiumicino, Italy, or a battleship with the same name. By linking the text with the correct URI, it becomes clear what the text refers to. Secondly, linked data allows us to retrieve additional information about a certain entity in an automated way, build connections between different resources and contextualize them [9]. For example, it allows us to link items tagged with the term “ring” with the broader concept of “jewelry” and, thus, interconnect them with items enriched with the term “bracelet”, which is also an instance of ”jewelry ”. Moreover, linked data usually comes with translated labels, thus improving the capabilities for multilingual search [10, 12]:

Semantic enrichment is a labour-intensive process, which requires effort and resources that CH institutions often lack. State-of-the-art AI technologies can be employed to automate the time-consuming and often mundane process of manual metadata enrichment. Natural language processing (NLP) tools can be used to analyse textual metadata and detect and classify concepts or named entities mentioned in unstructured text. Machine Learning (ML) approaches are extensively used for the task of disambiguation, which is responsible for deciding if the reference to ‘Leonardo da Vinci’ in the text refers to the Italian polymath or to the battleship. However, the accuracy of the automatic results highly hinges on the specific task at hand vis-a-vis the algorithm applied. For example, short textual descriptions, which are common in CH metadata, lack context and thus ML algorithms trained on Wikipedia articles may result in many incorrect matches. For similar reasons, they may often miss domain-specific matches that are relevant in the specific CH context. What’s more, even if the automatically detected links are correct, they may be considered undesirable for a certain case study. For example, linking metadata records with terms representing colours may be important for a fashion collection, but it may be undesirable for describing a manuscript that happens to mention a certain colour.

As a result, depending on a number of factors, such as the text characteristics (e.g. its length and language), the vocabulary that we wish to link it to, and the type of entities to detect (e.g. do we wish to identify a broad variety of concepts or to limit ourselves to certain domain-specific terms?), a different combination of tools and steps is required to achieve the best possible results for each specific task. For example, for certain tasks with a well-defined restricted context, a simple lemmatisation and string matching approach may be more appropriate than complex

[4] https://www.europeana.eu

ML-based algorithms. Besides the need for flexibility in combining and experimenting with different approaches and tools, another crucial aspect that needs to be considered is the need to make human inspection and validation an integral part of the end-to-end semantic enrichment workflow [13]. Given that manual validation is a resource-consuming task, practically, evaluation focuses on an appropriately selected sample of all the automatic annotations, depending on the collected feedback and the objective, appropriate filtering criteria are applied.

To address the aforementioned challenges, in this paper, we define, implement, and test a methodology and associated digital platform, called SAGE[5], which combines automatic annotation tools with human validation for the enrichment of CH items at scale. SAGE is an open source tool[6] that streamlines and facilitates the whole workflow of semantic enrichment, from data import and the automatic production of semantic annotations to human validation and data publication. The platform has been configured to serve the needs of the cultural sector and supports seamless interoperability with the common European data space for CH[7] and in particular with Europeana.

The methodology and platform have been applied to enrich the metadata records from datasets on various aspects of crafts heritage (from furniture to jewelry and costumes to clocks) coming from 8 different CH organisations, including the Fashion Museum Antwerp, the Netherlands Institute for Sound and Vision, the Open University of the Netherlands, the Greek National Documentation Centre, the Museum of Arts and Crafts in Zagreb, the Palais Galliera and Mobilier National in Paris, and the Textile Museum of Prato. The rest of the paper is structured as follows. After discussing related work, we present the steps of the methodology to semantic enrichment that we followed along with the technical architecture and the supporting SAGE platform, the evaluation performed and the results achieved. Finally, we conclude the paper with some general lessons learned.

2. Related Work

State-of-the-art Natural Language Processing and Machine Learning technologies have been extensively used in the CH domain to analyze unstructured text and extract structured information from it. To achieve automated subject indexing, Annif [22] is an open-source multilingual toolkit by the National Library of Finland that automatically assigns documents with subjects from a controlled vocabulary. In [1], a topic detection approach is applied to group historical documents into thematic collections. Additionally, the HerCulB system [23] has been developed to automatically annotate the Balkans' intangible CH. Other approaches propose the use of semi-automatic tools to assist humans in the task of manual annotation by identifying alignments between vocabularies, such as CultuurLINK [16].

Among information retrieval approaches, there have been several attempts to apply Named Entity Recognition (NER) as well as Disambiguation (NED) in the CH and digital humanities sectors, considering different types of data. In [11], NERD is applied to enrich metadata for the

[5] https://pro.europeana.eu/post/close-encounters-with-ai-an-interview-on-automatic-semantic-enrichment

[6] Source code: https://github.com/ails-lab/sage-backend and https://github.com/ails-lab/sage-frontend
Documentation: https://ails-lab.github.io/SAGE_Documentation/ and https://www.youtube.com/playlist?list=PLZhh656xkjIsxMKShH7aV7aR8TAwmU508

[7] https://dataspace-culturalheritage.eu/

exhibits of the Smithsonian Cooper–Hewitt National Design Museum in New York. In [8], an overview of NER approaches applied to historical documents is provided. An entity matching approach that works at the level of structured knowledge graphs, aiming to identify duplicate entities in data sources containing historical data is presented in [2]. In [3], the authors conduct a comparative study of different NERD tools on digital archive collections in order to link Engish textual metadata to Wikidata entities. In their study, the multilingual NERD tool mGENRE [6], which we employ in the current study, outperforms other approaches including BLINK [24] and EDGEL [14]. The need to deal with multilingual text is another important concern in the CH domain, e.g. named entity recommendation has been explored as a means to enhance multilingual retrieval on Europeana [10]. In this respect, the multilingual autoregressive entity linking approach employed by mGENRE is another advantage of the particular tool.

It should also be noted that NERD tools are trained on generic corpora [6, 24], that have limited overlap with CH-related textual metadata [12]. Adapting these tools to new domains by fine tuning them requires large amounts of well-annotated data, with labels that need to be generated or validated by domain experts, as well as large computational power, time and funds. These challenges are extensively discussed in [21] for the domain of Digital Humanities. Although domain adaptation of ML models is beyond the scope of the current paper, the methodology we advocate can lead to the production of high-quality ground truth data with reduced costs: validators are provided with datasets that have been already automatically annotated, an approach that highly facilitates their manual task, which becomes more focused and less cumbersome. This process allows us to make openly available a selection of appropriately processed annotated metadata from the CH domain (see Section 4), thus contributing to increasing the availability of annotated metadata that can be used for the in-domain tuning of NERD tools.

As the uptake of AI tools is expanding, there is increasing need for validation and moderation by humans to overcome the errors of the machine and achieve higher quality results [20]. Crowdsourcing methods and tools have been employed by CH organisations in this respect [13] as a means to mobilise human participants in the evaluation and correction of AI algorithm outcomes, also leading to the preparation of ground-truth data [15, 12]. For tasks that require specialised expertise, in [7] a niche-sourcing methodology and tool for the annotation of CH metadata is proposed, which, similar to our approach, uses an RDF triple store to store the results. However, as opposed to the the current work, the methodology relies solely on manual selections by experts with no use of automatic annotation tools.

Overall, our work distinguishes itself from previous work on semantic enrichment mainly in that it is based on a generic data management approach, which allows the combination of various annotation tools with flexible parameterisation capabilities (such as the definition of string matching and filtering rules); in that it includes human validation as an integral part of its workflow; and that it supports integrations with other CH-specific data representations and platforms, making it readily reusable in the CH data space. It should be noted that the integration with external annotation tools and CH-related platforms is loosely coupled, via interactions with the APIs (Application Programming Interface) and SPARQL endpoints exposed by the third-party components.

3. Methodology and Technical Architecture

The methodology we followed for the semantic enrichment of CH metadata consists of the following high-level steps:

1. *A: Data aggregation and requirements analysis*
   The first step concerns the preparatory tasks of aggregating the data and specifying the requirements for the the enrichment (e.g. which metadata fields to analyse, which vocabularies to link to etc).
2. *B: Automatic metadata enrichment*
   The second step involves the automatic analysis of the textual metadata, with the aim to derive useful annotations in line with the identified requirements.
3. *C: Human validation*
   Humans are solicited to review and validate the automatically generated annotations as well as to manually add new annotations, that the automatic algorithm has not been able to detect.
4. *D: Filtering and data publication*
   The outcomes of the human validation are analysed to establish appropriate thresholds for filtering and the filtered annotations are embedded as enrichments to the metadata records. The enriched metadata records are ultimately published to the Europeana platform.

Figure 1 provides an overview of the main digital components that support the above methodology.

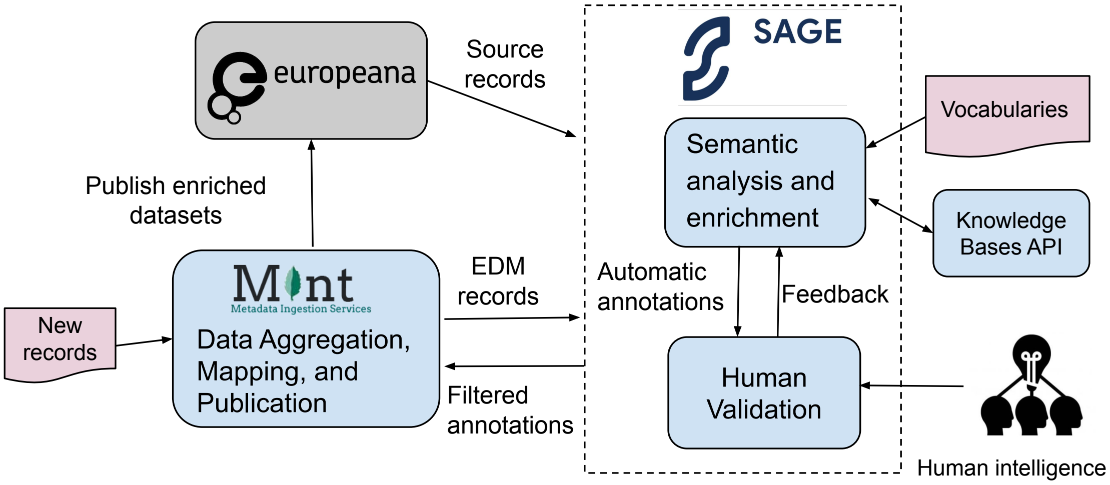


Figure 1: Architectural Overview

MINT is a metadata management tool[8] that is part of the data space for CH and is used by several aggregators to prepare and publish their data to Europeana. It acts as the link between

[8] https://mint-wordpress.image.ntua.gr/

SAGE and Europeana and supports steps A and D of the aforementioned methodology, serving the following purposes: (i) aggregate the metadata records from the data providers and for mapping them to the Europeana Data Model (EDM) [4] that is then passed to SAGE; and (ii) embed the annotations produced by SAGE, after filtering in light of the human feedback, into the original metadata records in line with the expected EDM extension that accommodates for enrichments[9] and ultimately publishing the results to Europeana. It should be noted that data already published on the Europeana platform can also be sourced directly by SAGE for annotation, via a direct interconnection with the Europeana search API[10].

3.1. The SAGE tool for automatic enrichment and validation

SAGE is a web-based platform for generating, enriching, validating, publishing, and searching RDF data. In the context of our methodology, it is responsible for the core steps B and C. The RDF data can be produced from heterogeneous data sources and data formats using the D2RML mapping language [5], and enriched using annotators that wrap web-based or other third party services. The enrichments can then be manually validated, and finally, the entire data can be published in an RDF store and indexed. The SAGE platform has been configured to facilitate the semantic enrichment of CH metadata. In this respect, it offers a suite of already set-up annotators, i.e. parameterisable enrichment templates, that are connected with relevant in-domain vocabularies and knowledge bases. It also facilitates the direct import/publication of metadata from/to platforms of the European data space for CH, including Europeana and MINT, making use of established APIs and formats .

A dataset is annotated per property, i.e. the user can select from the schema preview a property that links entities to values, and execute an annotator on the values of that property. An annotator in SAGE is a mediator that retrieves all desired values from the triple store where the dataset content is published, generates the appropriate calls to the web or other service, and transforms the results to the RDF annotation specification. As in the case of datasets, the results of an annotator execution are Terse RDF Triple Language[11] files stored in the file system of SAGE. In the framework of the data space of CH, annotations are also expressed in a JSON-LD equivalent representation model[12], which bases on the W3C's Web Annotation Model[13] supported by Europeana. The annotation model is generic enough to accommodate for various enrichment types (e.g. annotations resulting from automatic translation tools, from image analysis etc) and provides sufficient provenance information, including information about the annotations' confidence scores and the validation feedback provided by humans. For metadata records that are compatible with EDM, the annotations are ultimately embedded in the metadata in line with the EDM extension that instructs the representation of metadata statements resulting from semantic enrichment[14]. This way, the enrichments can be appropriately

[9] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_profiles/EDM_provenance_profile_external_202111.pdf

[10] https://www.europeana.eu/en/apis

[11] https://www.w3.org/TR/turtle/

[12] https://docs.google.com/document/d/1Cq1Qqx0ji7Vw8iwLVis1CfpYKtv-72ojkcvjnQzrKjs/edit?usp=sharing

[13] https://www.w3.org/TR/annotation-model/

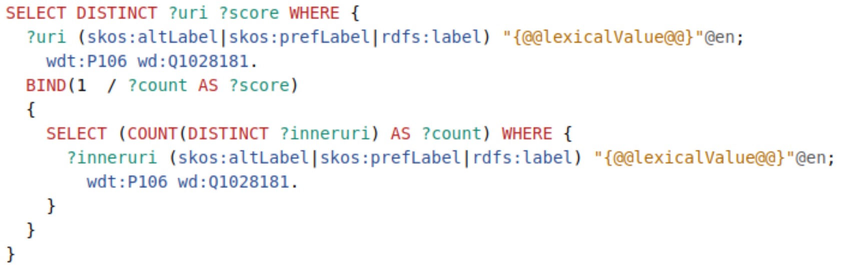
[14] https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_profiles/EDM_provenance_profile_external_202111.pdf

handled and presented to the end-user by the Europeana platform.

SAGE supports three main types of annotators, which can be parameterised with respect to different aspects (e.g. vocabulary, language, preprocessing functions etc) to serve different case studies:

- Thesaurus annotators: They link texts to URIs from thesauri that can be imported to the platform by performing smart string matching on the thesaurus labels using lemmatizers (such as the ones provided by the Stanza library[15]) and other functions to produce improved results (e.g. apply dedicated Regex rules). They are appropriate for application both on generic textual fields and on focused short fields. By selecting thesauri that represent concepts referring to specific domains (e.g. fashion), it is more likely that the extracted terms are relevant to the object in question. Moreover, such annotators can perform massive enrichments in a very short time compared to the other annotators, since they rely on locally stored data. Figure 3 provides an overview of how a Thesaurus Annotator works on a specific example.
- Generic NERD annotators: They employ pre-trained NERD tools to detect named entities and link them to respective entities from Wikidata. SAGE supports two different pipelines for generic NERD. The first pipeline makes use of the AIDA tool [18] for entity detection and disambiguation. The second pipeline makes use of the spaCy library[16] for performing the NER part for different languages, i.e. for recognising entities and their string boundaries within a sentence, and then of the multilingual mGENRE model [6] for the disambiguation stage and for linking with a URI from Wikidata. Such annotators can be used as they are, with minimal or no configurations and are appropriate for general-purpose enrichments. They conduct disambiguation by using the context contained in longer texts (e.g. description), since they are trained on textual corpora such as Wikipedia articles. At the same time, this process is more likely than the other annotators to link with terms that are too generic or irrelevant in the context of a specific case study, while it is hard to infer with sufficient accuracy the type of the extracted entity and its relation to the object in question (e.g. whether it represents the item's creator, a place of display etc). As a result, in practice, they often produce more accurate results when applied in fields with pre-specified focused semantics.
- SPARQL Annotators: SPARQL annotators communicate with external knowledge bases (such as Wikidata and Geonames) through SPARQL endpoints. Thus, they are the best fit when dealing with large knowledge bases that cannot be downloaded locally.They can be applied on focused fields that refer to a single entity. The values of such fields often follow certain patterns (e.g. "surname, name", "city/region/country" etc) and, thus, pre-processing with Regex is key to the success of the method, so that a normal form of the entity name can be extracted. An example of a query that matches Wikidata entities with the occupation of a painter is presented in Figure 2.

[15] https://stanfordnlp.github.io/stanza/
[16] https://spacy.io/



```
SELECT DISTINCT ?uri ?score WHERE {
  ?uri (skos:altLabel|skos:prefLabel|rdfs:label) "{@@lexicalValue@@}"@en;
    wdt:P106 wd:Q1028181.
  BIND(1  / ?count AS ?score)
  {
    SELECT (COUNT(DISTINCT ?inneruri) AS ?count) WHERE {
      ?inneruri (skos:altLabel|skos:prefLabel|rdfs:label) "{@@lexicalValue@@}"@en;
        wdt:P106 wd:Q1028181.
    }
  }
}
```

Figure 2: An example of a SPARQL query searching Wikidata. It matches labels with English Wikidata labels of items having an occupation (wdt:P106) painter (wd:Q1028181). It also estimates a confidence score as $1/(number\_of\_matches)$.

3.2. Human Validation

Human validation was conducted via a dedicated environment provided by SAGE (see Figure 4). Humans are invited to inspect the automatic annotations produced by the AI tools and accept or reject them. Moreover, they can add missed annotations, i.e. relevant annotations that the automatic algorithm failed to identify. During the validation of the results of the semantic analysis, validators are also able to edit the predefined target metadata field in which the URI will end up. It should be noted that SAGE groups together annotations repeated across many records in a dataset and flags annotations referring to URIs that are already included in the metadata. In total, 14 CH professionals with specialized knowledge about the considered collections participated in the validation process, with two to three validators per collection. Participants were instructed to accept or reject annotations based on what they consider as desirable for inclusion in the final metadata. That is, they evaluated not only whether an annotation is a correct match but also in terms of relevance (e.g. matches with the term ”human”

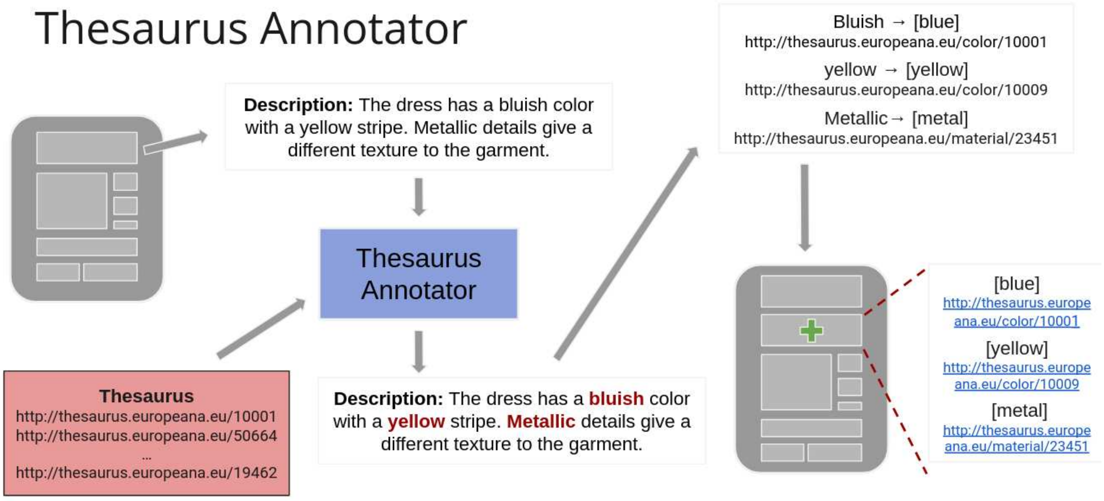
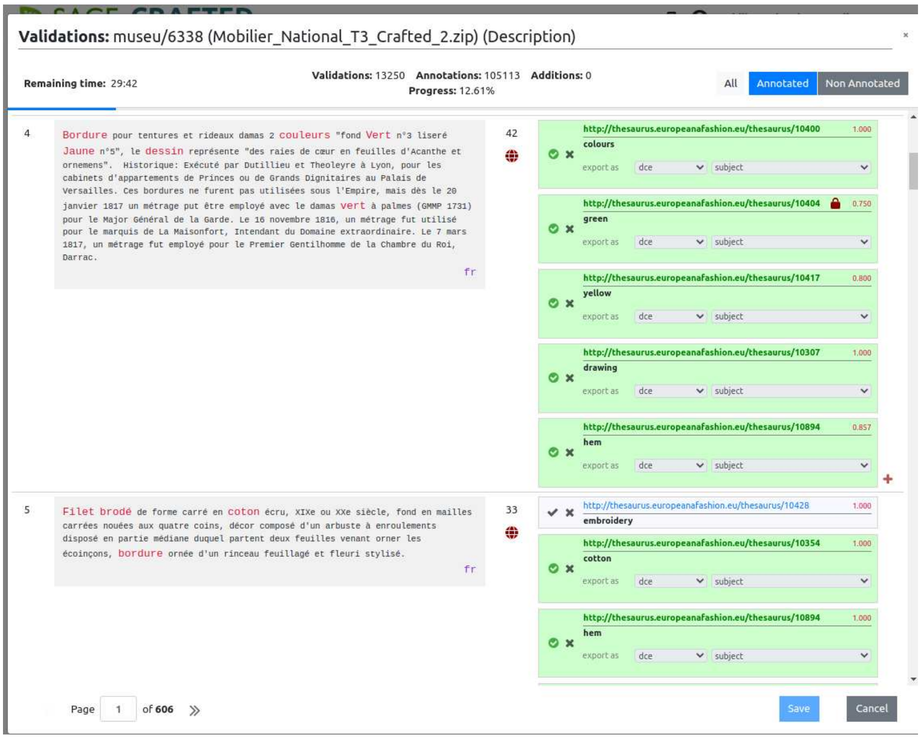

Thesaurus Annotator
Description: The dress has a bluish color with a yellow stripe. Metallic details give a different texture to the garment.
Bluish → [blue]
http://thesaurus.europeana.eu/color/10001
yellow → [yellow]
http://thesaurus.europeana.eu/color/10009
Metallic→ [metal]
http://thesaurus.europeana.eu/material/23451
Thesaurus Annotator
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.
[blue]
http://thesaurus.europeana.eu/color/10001
[yellow]
http://thesaurus.europeana.eu/color/10009
[metal]
http://thesaurus.europeana.eu/material/23451

Figure 3: Overview of the SAGE Thesaurus Annotator workflow on a metadata description.

may be considered too generic) .

The appropriate size and characteristics of the sample to be validated depend on the available resources that can be invested in the validation process and the nature of the use case. What is considered a ”sufficient” amount hinges on many factors, including the total number of automatically produced annotations, their characteristics (e.g. what metadata fields they refer to, their granularity, etc), the characteristics of the automated algorithm that produced them (e.g. its accuracy, the reliability of the automatic confidence scores it assigned to them), the number of participants and the amount of time they can devote to the task. The following criteria were used to guide the selection of the annotations sample to be validated, so as to ensure representativeness across various parameters:

- Inspect annotations that appear in a high number of records and thus will have a high impact.
- Ensure a balanced representation of metadata fields, including fields with varying semantics and expected text length.
- Take into consideration automatic confidence levels assigned by automatic algorithms, if available: inspect annotations with a rather low confidence score but also a sufficient number of annotations with a rather high one.



Validations: museu/6338 (Mobilier_National_T3_Crafted_2.zip) (Description)

Remaining time: 29:42

Validations: 13250 **Annotations:** 105113 **Additions:** 0
Progress: 12.61%

4 Bordure pour tentures et rideaux damas 2 couleurs "fond Vert n°3 liseré Jaune n°5", le dessin représente "des raies de cœur en feuilles d'Acanthe et ornemens". Historique: Exécuté par Dutillieu et Theoleyre à Lyon, pour les cabinets d'appartements de Princes ou de Grands Dignitaires au Palais de Versailles. Ces bordures ne furent pas utilisées sous l'Empire, mais dès le 20 janvier 1817 un métrage put être employé avec le damas vert à palmes (GMMP 1731) pour le Major Général de la Garde. Le 16 novembre 1816, un métrage fut utilisé pour le marquis de La Maisonfort, Intendant du Domaine extraordinaire. Le 7 mars 1817, un métrage fut employé pour le Premier Gentilhomme de la Chambre du Roi, Darrac. fr

42

URI	Label	Score	export as
http://thesaurus.europeanafashion.eu/thesaurus/10400	colours	1.000	dce subject
http://thesaurus.europeanafashion.eu/thesaurus/10404	green	0.750	dce subject
http://thesaurus.europeanafashion.eu/thesaurus/10417	yellow	0.800	dce subject
http://thesaurus.europeanafashion.eu/thesaurus/10307	drawing	1.000	dce subject
http://thesaurus.europeanafashion.eu/thesaurus/10894	hem	0.857	dce subject

5 Filet brodé de forme carré en coton écru, XIXe ou XXe siècle, fond en mailles carrées nouées aux quatre coins, décor composé d'un arbuste à enroulements disposé en partie médiane duquel partent deux feuilles venant orner les écoinçons, bordure ornée d'un rinceau feuillagé et fleuri stylisé. fr

33

URI	Label	Score	export as
http://thesaurus.europeanafashion.eu/thesaurus/10428	embroidery	1.000	
http://thesaurus.europeanafashion.eu/thesaurus/10354	cotton	1.000	dce subject
http://thesaurus.europeanafashion.eu/thesaurus/10894	hem	1.000	dce subject

Page 1 of 606

Figure 4: Screenshot from the SAGE validation environment. Strings of metadata that have been matched are shown on the left and the URI(s) they have been linked to on the right.

3.3. Analysis and Filtering of Annotations

Validation feedback was analysed with the aim of establishing thresholds for annotations that are considered acceptable for publication. To this end, the following metrics have been calculated per dataset, per Annotator, and per analysed metadata field:

- Precision considering only unique annotations, that is, unique triples of field textual values, matched sub-string, and identified URI.
- Precision considering all annotations, that is, without grouping together identical field textual values (in other words, counting all times the same annotation, defined as a triple, may appear in different items).

In both cases, precision was calculated as $TP/(TP + FP)$, where $TP = \textit{accepted annotations}$ and $FP = \textit{rejected annotations}$. Precision was used as a threshold for filtering out not-reviewed annotations on a field or Annotator basis. What is considered a sufficiently high precision depends on the requirements of each case study and the expectations of the data provider.

For the use cases we considered, most human experts did not focus on the manual insertion of new annotations and the few manually added annotations we collected do not allow us to sufficiently estimate false negatives and thus compute recall and the F-score. It should also be noted that for publication to Europeana, a threshold based on precision is considered the most appropriate metric to be used[17].

Human judgments can also be used as a means to assess the trustworthiness of the automatic confidence scores assigned by the AI algorithms. For example, if humans tend to accept all sample annotations above a certain score, then we may conclude that all annotations above that score can be regarded as acceptable. In this vein, we explored whether there is a correlation between the automatic confidence scores, when available, and human judgments. We therefore plotted the logistic regression between the two variables considering the following metrics:

- The $p - value$ [19]. A value greater than 0.05 means that no statistically significant relationship between the automatic scores and the human judgments was observed.
- The expected automatic score for which the predicted probability is greater than 0.7, that is annotations above this score have a probability above 0.7 to be accepted by humans, based on the sample data.

4. Results on Crafts Heritage Datasets

The aforementioned methodology and supporting tools have been applied to metadata records describing crafts heritage items as mentioned in Section 1. The analysed metadata comes in the following languages: Dutch, Italian, French, Greek, English and Croatian. In total, the SAGE annotators were applied on 216, 115 metadata records, giving rise to 915, 472 total annotations and 549, 402 unique annotations. It should be noted that numbers calculated based on unique annotations are considered more reliable since numbers that count in item impact are skewed towards textual values that are repeated in multiple items. In total, 12 experts from 8 CH organisations took part in the validation campaigns. Overall, 30, 910 unique annotations referring to

[17] https://pro.europeana.eu/post/methodology-for-validating-enrichments

more than 15K records were reviewed via SAGE (i.e. 5.6% of the automatically produced unique annotations), with the sample being selected following the criteria outlined in Section 3.2. Of those annotations, 23,426 were accepted and 7,474 were rejected.

The overall precision, defined as the number of all accepted automatic annotations produced by SAGE over the number of reviewed automatic annotations, is 0.76, considering unique annotations. If all annotations are counted in, then the overall precision is 0.82. Precision varied largely depending on the analysed metadata field, the type of annotator that was used, and the datasets that were analysed. Table 1 provides an overview of the results achieved by different annotators. The minimum and maximum precision reported in the table refer to a per metadata field level.

Table 1
Precision of used SAGE annotators

Annotator	Min Precision (all/unique)	Max Precision (all/unique)	Avg precision / (all/unique)
Fashion Thesaurus Annotator	0.436/0.672	0.964/0.943	0.801/0.832
AAT Annotator	0.644/0.658	0.994/0.987	0.819/0.822 .
Greek Crafts Thesaurus Annotator	0.982/0.947	0.982/0.947	0.982/0.947
EUScreen Thesaurus Annotator	0.607/0.878	0.952/0.927	0.779/0.902
Wikidata SPARQL	0.894/0.817	1/1	0.981/0.963
Generic NERD with Wikidata - mGENRE	0.4/0.4	1/1	0.935/0.748

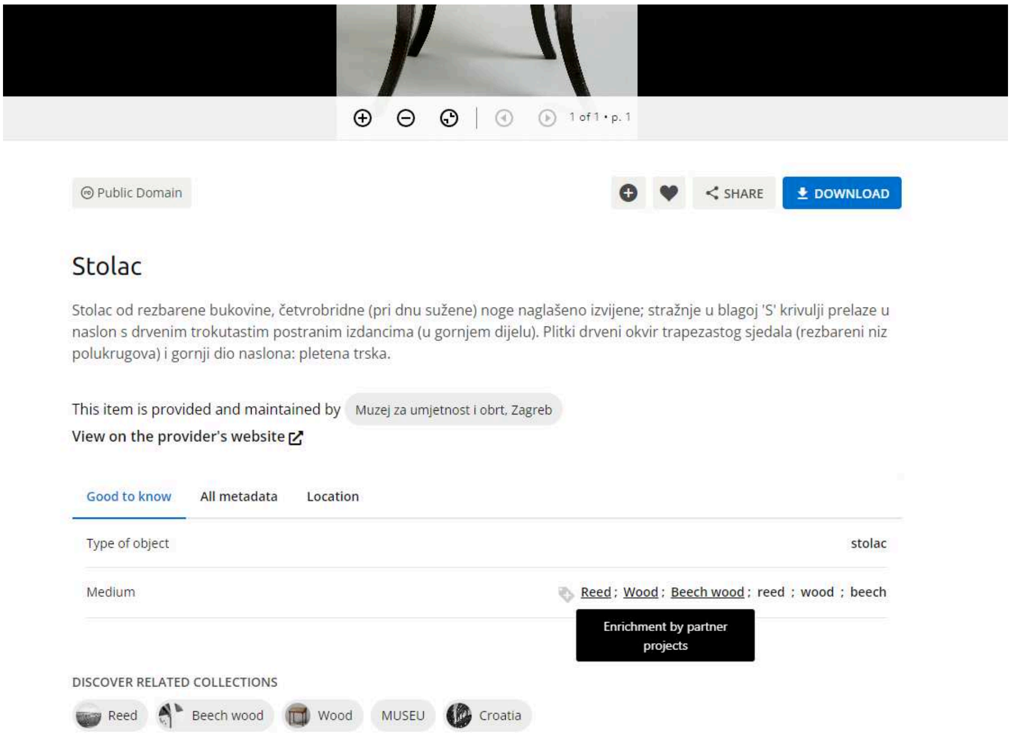
The choice of the vocabulary used by the thesaurus annotators depended on the respective dataset characteristics and providers' objectives. The following vocabularies were used: the Europeana fashion thesaurus[18]; AAT; the EUScreen vocabulary on audiovisual heritage[19]; and a SKOS vocabulary on Greek crafts heritage[20]. Thesaurus annotators were applied to both longer (e.g. `dc:description`, `dc:title`) and shorter fields (e.g. `dc:format`, `dc:type`), often after case-appropriate regex pre-processing, giving rise to generally satisfactory results in both cases. SPARQL queries on Wikidata were used to retrieve creators for the `dc:creator` and locations for the `dc:spatial` fields. Although in most cases it did not produce a high number of annotations, it scored a high precision. mGENRE and AIDA were applied to `dc:description` and `dc:title` fields as well as shorter fields (including `dc:creator`, `dc:spatial`, and `dc:rights`). They both produced similar results, performing well for short fields but poorly for longer ones. In the latter case, they both struggled with disambiguation between multiple candidate entities and, even when producing matches that were in principle correct, those were often too generic and considered irrelevant by validators.

For annotators for which an automatic score was produced, we also attempted to plot the logistic regression between the automatic score and the human judgments. However, no correlation was found between the two variables and therefore automatic scores were not used as factors in the filtering rules. A possible explanation for this is that in the case of thesauri

[18] http://thesaurus.europeanafashion.eu/

[19] http://thesaurus.euscreen.eu/EUscreenXL/v1

[20] https://www.semantics.gr/authorities/vocabularies/craft-item-types/vocabulary-entries



Stolac

Stolac od rezbarene bukovine, četvrobridne (pri dnu sužene) noge naglašeno izvijene; stražnje u blagoj 'S' krivulji prelaze u naslon s drvenim trokutastim postranim izdancima (u gornjem dijelu). Plitki drveni okvir trapezastog sjedala (rezbareni niz polukrugova) i gornji dio naslona: pletena trska.

This item is provided and maintained by Muzej za umjetnost i obrt, Zagreb

View on the provider's website

Type of object	stolac
Medium	Reed ; Wood ; Beech wood ; reed ; wood ; beech

Enrichment by partner projects

Figure 5: View on Europeana of an item provided by the Museum of Arts and Crafts in Zagreb. The 'Reed', 'Wood', and 'Beech Wood' terms are all automatic enrichments added by SAGE that are visible on the item page.

annotators, scores are usually quite high for all annotations: they reflect the string difference (1-Levenshtein distance [17]) between words endings (since the matching is based on the lemmatised versions of the textual metadata and the thesaurus terms). For the generic NERD tools the scores turn to be quite unreliable: they are inversely proportional to the number of candidate URIs and do not sufficiently account for disambiguation.

Annotations have been filtered by discarding all annotations rejected by humans, while including all explicitly accepted ones. (considering a majority vote). For non-reviewed annotations, a threshold based on precision between 0.75 and 0.8 (considering unique annotations) was considered acceptable by data providers. In total, 549.460 have been regarded acceptable, leading to the enrichment of 133.405 out of 216.115 analysed records. All enriched records have been published to Europeana. Enrichments have been indexed to become searchable and are visible as part of the item view via distinct tags, thus contributing to making the respective items more discoverable, contextual, and multilingual. Figure 5 shows an example of how automatic annotations look like on the Europeana platform.

Although domain adaptation is beyond the scope of the current case study, the dataset that resulted from the validation process can be valuable for the training and fine-tuning of NERD tools in the field of CH. To this end, a curated selection of annotated metadata enriched and validated via SAGE has been made openly available[21] under a CC0 license, so that it can be freely reused as data amenable for computational purposes. The dataset includes more than 10K unique annotations (pairs of analysed textual values and URIs). The in-domain adaptation of NERD tools so that they can more effectively deal with the particular characteristics of CH metadata [12], such as short text and specialised terminology, remains part of future work.

5. Conclusions

In the current paper, we present a generic and reusable methodology and supporting digital platform that combines automatic annotators with human expertise in order to enrich them with terms from various linked data sources. The methodology has been applied and evaluated on a case study involving crafts heritage datasets, leading to measurable improvements in the quality of metadata and enhancing the discoverability and usability of the respective resources on Europeana. Building on the practical experience we gained, the current case study allows us to draw some lessons learned, which can prove useful for interested stakeholders who may wish to follow a similar process to enrich their datasets.

Before proceeding to the actual enrichment, it is crucial to scrutinise the data to be analysed, gain a deep understanding of its characteristics and define feasible and meaningful enrichment objectives. One should define the expected benefit of possible enrichments and how they will bring value to the collection. In this respect, one should ask questions such as: What kind of concepts are useful to detect (e.g. persons, locations, domain-specific concepts etc)? Which metadata fields contain relevant information (e.g. descriptions make frequent references to techniques and materials used)? In what languages are the metadata? It should also be noted that the quality of the original metadata affects the quality of the automatic enrichment. If the text contains many typos or is misaligned with the intended semantics of the respective metadata field, then the outputs of the automatic enrichment tools will be less accurate. This step is also crucial for detecting patterns in data that can be exploited in order to produce annotations.

The next step involves the selection and set-up of the semantic annotators that are most appropriate for the specific use case, considering the advantages and disadvantages of each approach as presented in Section 3.1. The selection of knowledge bases and vocabularies that have the case-appropriate granularity and coverage is crucial. Generally, the more focused the automatic enrichment, considering the terminology used (e.g. link with a domain-specific vocabulary versus general-purpose NERD) and the metadata property that is parsed (e.g. topic-specific fields such as `dc:creator` versus longer ones such as `dc:description`), the less the risk of producing too many irrelevant or too generic enrichments and the more accurate the resolution of disambiguation. One should opt for knowledge bases that are accessible on the Web via an open license, well-documented, and compliant with Linked Data best practices. Their multilingual coverage (also in relation to the language of your metadata) is also an important aspect that should be taken into consideration.

After the production of the automatic annotations, the validation process should be carefully

[21] See https://github.com/ails-lab/ai4culture-datasets for the actual dataset and the process that was used for the data curation.

organised. The background of the validators is crucial: some tasks may require expert skills (e.g., knowledge of a particular language, domain expertise etc.), while others can be performed by appealing to a general audience. In the former case, it is wiser to keep the validation process closed within a team of experts, while in the latter, organizing an open crowdsourcing campaign will mobilize more people and thus speed up the process. The selection of the sample to be validated is crucial: it does not need to be large but it should be well-balanced, following the criteria outlined in Section 3.2.

The final step involves the filtering of the automatic validation in light of the acquired human feedback. For annotations reviewed by humans, majority vote can typically be used to determine acceptability. Depending on the annotation type, additional criteria might be enforced (e.g. for public validation campaigns where untrustworthy feedback is suspected, we may require that an annotation is reviewed by multiple users). Automatic annotations that have not been reviewed by humans or lack a reliable confidence score should be filtered using automatic evaluation metrics. The appropriate metrics depend on the nature of the task, but precision is a typical one when correctness is at high stake. Thresholds should be established depending on what is considered acceptable given the specific use case requirements.

Acknowledgments

The work is co-funded by the European Union, under the projects “CRAFTED: Enrich and promote traditional and contemporary crafts” and “AI4Culture: An AI platform for the cultural heritage data space”. We would like to thank all partners of the CRAFTED project, and particularly Panagiotis Tzortzis, for their valuable contributions to this work.

References

[1] M. Andresel, S. Gordea, S. Stevanetic, and M. Schütz. “An Approach for Curating Collections of Historical Documents with the Use of Topic Detection Technologies”. In: *Int. J. Digit. Curation* 17.1 (2022), p. 12.

[2] J. Baas, M. M. Dastani, and A. Feelders. “Entity Matching in Digital Humanities Knowledge Graphs”. In: *Proc. of the Conf. on Computational Humanities Research, CHR2021.* Vol. 2989. CEUR Workshop Proceedings. 2021, pp. 1–15.

[3] Y. Benkhedda, A. Skapars, V. Schlegel, G. Nenadic, and R. Batista-Navarro. “Enriching the Metadata of Community-Generated Digital Content through Entity Linking: An Evaluative Comparison of State-of-the-Art Models”. In: *Proc. of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.* St. Julians, Malta: Association for Computational Linguistics, 2024, pp. 213–220.

[4] V. Charles, A. Isaac, V. Tzouvaras, and S. Hennicke. “Mapping Cross-Domain Metadata to the Europeana Data Model (EDM)”. In: *Research and Advanced Technology for Digital Libraries.* Springer Berlin Heidelberg, 2013, pp. 484–485.

[5] A. Chortaras and G. Stamou. "D2RML: Integrating Heterogeneous Data and Web Services into Custom RDF Graphs". In: *Workshop on Linked Data on the Web co-located with The Web Conference*. Vol. 2073. CEUR Workshop Proceedings. 2018.

[6] N. De Cao, L. Wu, K. Popat, M. Artetxe, N. Goyal, M. Plekhanov, L. Zettlemoyer, N. Cancedda, S. Riedel, and F. Petroni. "Multilingual Autoregressive Entity Linking". In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 274–290.

[7] C. Dijkshoorn, V. de Boer, L. Aroyo, and G. Schreiber. "Accurator: Nichesourcing for Cultural Heritage". In: *Hum. Comput.* 6 (2019), pp. 12–41.

[8] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet. "Named Entity Recognition and Classification in Historical Documents: A Survey". In: *ACM Computing Surveys* 56.2 (2023).

[9] N. Freire and A. Isaac. "Technical Usability of Wikidata's Linked Data". In: *Business Information Systems Workshops*. Ed. by W. Abramowicz and R. Corchuelo. Springer International Publishing, 2019, pp. 556–567.

[10] S. Gordea, M. L. Paramita, and A. Isaac. "Named Entity Recommendations to Enhance Multilingual Retrieval in Europeana.eu". In: *Foundations of Intelligent Systems*. Springer International Publishing, 2020, pp. 102–112.

[11] S. Hooland, M. Wilde, R. Verborgh, T. Steiner, and R. Van de Walle. "Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections". In: *Literary and Linguistic Computing* (2013).

[12] E. Kaldeli, M. García-Martínez, A. Isaac, P. S. Scalia, A. Stabenau, I. L. Almor, C. G. Lacal, M. B. Ordóñez, A. Estela, and M. Herranz. "Europeana Translate: Providing multilingual access to digital cultural heritage". In: *Proc. of the 23rd Annual Conference of the European Association for Machine Translation, EAMT*. European Association for Machine Translation, 2022, pp. 297–298.

[13] E. Kaldeli, O. Menis-Mastromichalakis, S. Bekiaris, M. Ralli, V. Tzouvaras, and G. Stamou. "CrowdHeritage: Crowdsourcing for Improving the Quality of Cultural Heritage Metadata". In: *Information* 12.2 (2021).

[14] N. Lai. "LMN at SemEval-2022 Task 11: A Transformer-based System for English Named Entity Recognition". In: *Proc. of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 1438–1443.

[15] V. Lyberatos, S. Kantarelis, E. Kaldeli, S. Bekiaris, P. Tzortzis, O. Menis - Mastromichalakis, and G. Stamou. "Employing Crowdsourcing for Enriching a Music Knowledge Base in Higher Education". In: *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*. Springer Nature, 2023, pp. 224–240.

[16] H. Manguinhas, V. Charles, A. Isaac, T. Miles, A. Lima, A. Neroulidis, V. Ginouvès, D. Atsidis, M. Hildebrand, M. Brinkerink, and S. Gordea. “Linking Subject Labels in Cultural Heritage Metadata to MIMO Vocabulary using CultuurLink”. In: *Proc. of the 15th European Networked Knowledge Organization Systems Workshop (NKOS) co-located with the 20th Int. Conf. on Theory and Practice of Digital Libraries (TPDL).* Vol. 1676. CEUR Workshop Proceedings. 2016, pp. 32–35.

[17] F. P. Miller, A. F. Vandome, and J. McBrewster. *Levenshtein Distance: Information theory, Computer science, String (computer science), String metric, Damerau?Levenshtein distance, Spell checker, Hamming distance.* Alpha Press, 2009.

[18] D. B. Nguyen, J. Hoffart, M. Theobald, and G. Weikum. “AIDA-light: High-Throughput Named-Entity Disambiguation”. In: *Proc. of the Workshop on Linked Data on the Web co-located with the 23rd Int. World Wide Web Conf. (WWW.* Vol. 1184. CEUR Workshop Proceedings. 2014.

[19] S. Silvey. *Statistical Inference.* Monographs on statistics and applied probability. Chapman & Hall, 2003.

[20] J. Stiller, V. Petras, M. Gäde, and A. Isaac. “Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences”. In: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection.* Springer International Publishing, 2014, pp. 238–247.

[21] O. Suissa, A. Elmalech, and M. Zhitomirsky-Geffet. “Text analysis using deep neural networks in digital humanities and information science”. In: *Journal of the Association for Information Science and Technology* 73 (2021).

[22] O. Suominen, J. Inkinen, and M. Lehtinen. “Annif and Finto AI: Developing and Implementing Automated Subject Indexing”. In: *Italian Journal of Library, Archives and Information Science* 13.1 (2022), pp. 265–282.

[23] I. Tanasijević and G. Pavlović-Lažetić. “HerCulB: content-based information extraction and retrieval for cultural heritage of the Balkans”. In: *The electronic library* 38.5/6 (2020), pp. 905–918.

[24] L. Wu, F. Petroni, M. Josifoski, S. Riedel, and L. Zettlemoyer. “Scalable Zero-shot Entity Linking with Dense Entity Retrieval”. In: *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP).* 2020, pp. 6397–6407.

[25] M. Wu, H. Brandhorst, M.-C. Marinescu, J. M. Lopez, M. Hlava, and J. Busch. “Automated metadata annotation: What is and is not possible with machine learning”. In: *Data Intelligence* 5.1 (2023), pp. 122–138.