An Open Science System for Text Mining

Gianpaolo Coro
ISTI-CNR
via Moruzzi 1 Pisa, Italy
coro@isti.cnr.it

Giancarlo Panichi
ISTI-CNR
via Moruzzi 1 Pisa, Italy
panichi@isti.cnr.it

Pasquale Pagano
ISTI-CNR
via Moruzzi 1 Pisa, Italy
pagano@isti.cnr.it

Abstract

Text mining (TM) techniques can extract high-quality information from big data through complex system architectures. However, these techniques are usually difficult to discover, install, and combine. Further, modern approaches to Science (e.g. Open Science) introduce new requirements to guarantee reproducibility, repeatability, and re-usability of methods and results as well as their longevity and sustainability. In this paper, we present a distributed system (NLPHub) that publishes and combines several state-of-the-art text mining services for named entities, events, and keywords recognition. NLPHub makes the integrated methods compliant with Open Science requirements and manages heterogeneous access policies to the methods. In the paper, we assess the benefits and the performance of NLPHub on the I-CAB corpus[1].

1 Introduction

Today, text mining operates within the challenges introduced by big data and new Science paradigms, which impose to manage large volumes, high production rate, heterogeneous complexity, and unreliable content, while ensuring data and methods longevity through re-use in complex models and processes chains. Among the new paradigms, Open Science (OS) focusses on the implementation in computer systems of the three "R"s of the scientific method: Reproducibility, Repeatability, and Re-usability (Hey et al., 2009; EU Commission, 2016). The systems envisaged by OS, are based on Web services networks that support big data processing and the open publication of results. Although text mining techniques exist that can tackle big data experiments (Gandomi and Haider, 2015; Amado et al., 2018), few examples that incorporate OS concepts can be found (Linthicum, 2017). For example, common text mining "cloud" services do not allow easy repeatability of the experiments by different users and are usually domain-specific and thus poorly re-usable (Bontcheva and Derczynski, 2016; Adedugbe et al., 2018). Available multi-domain systems do not use communication standards (Bontcheva and Derczynski, 2016; Wei et al., 2016), and the few OS-oriented initiatives that use text mining focus specifically on documents preservation and cataloguing (OpenMinTeD, 2019; OpenAire, 2019).

In this paper, we present a multi-domain text mining system (*NLPHub*) that is compliant with OS and combines multiple and heterogeneous processes. NLPHub is based on an e-Infrastructure (e-I), i.e. a network of hardware and software resources that allow remote users and services to collaborate while supporting data-intensive Science through cloud computing (Pollock and Williams, 2010; Andronico et al., 2011). Currently, NLPHub integrates 30 state-of-the-art text mining services and methods to recognize fragments of a text (*annotations*) associated with named abstract or physical objects (named entities), spatiotemporal events, and keywords. These integrated processes cover overall 5 languages (English, Italian, German, French, and Spanish), requested by the European projects this software is involved in (i.e. (Parthenos, 2019; SoBigData, 2019; Ariadne, 2019)). These processes come from different providers that have different access policies, and the e-I is used both to manage this heterogeneity and to possibly speed up the processing through cloud computing. NLPHub uses the Web Processing Service standard (WPS, (Schut and Whiteside, 2007)) to describe all integrated processes, and the Prov-O XML

ontological standard (Lebo et al., 2013) to track the complete set of input, output, and parameters used for the computations (*provenance*). Overall, these features enable OS-compliance and we show that the orchestration mechanism implemented by NLPHub adds effectiveness and efficiency to the connected methods. The name "NLPHub" refers to the forthcoming extensions of this platform to other text mining methods (e.g. sentiment analysis and opinion mining), and natural language processing tasks (e.g. text-to-speech and speech processing).

2 Methods and tools

2.1 E-Infrastructure and Cloud Computing Platform

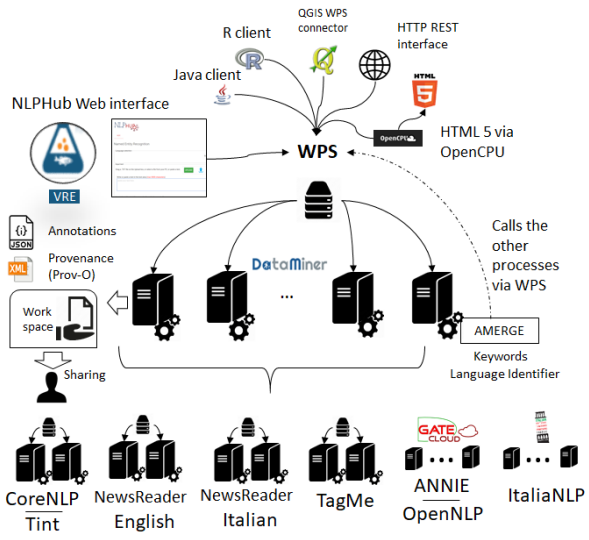

Figure 1: Overall architectural schema of the NLPHub.

NLPHub uses the open-source D4Science e-I (Candela et al., 2013; Assante et al., 2019), which currently supports applications in many domains through the integration of a distributed storage system, a cloud computing platform, online collaborative tools, and catalogues of metadata and geospatial data. D4Science supports the creation of Virtual Research Environments (VREs) (Assante et al., 2016), i.e. Web-based environments fostering collaboration and data sharing between users and managing heterogeneous data and services access policies. D4Science grants each user with access to a private online file system (the *Workspace*) that uses a high-availability distributed storage system behind the scenes, and enables folders creation and sharing between VRE users. Through VREs and accounting and security services, D4Science is able to manage heterogeneous access policies by granting free access to open services in public VREs, and controlled/private access to non-open services in private or moderated VREs. D4Science includes a cloud computing platform named DataMiner (Coro et al., 2015; Coro et al., 2017) that currently hosts ~400 processes and makes all integrated processes available under the WPS standard (Figure 1). WPS is supported by third-party software and allows standardising a process' input, its parameterisation and output. DataMiner executes the processes in a cloud computing cluster of 15 machines with Ubuntu 16.04.4 LTS x86 64 operating system, 16 virtual cores, 32 GB of RAM and 100 GB of disk space. These machines are hosted by the National Research Council of Italy and the Italian Network of the University and Research (GARR). Each process can parallelise an execution either across the machines (using a Map-Reduce approach) or on the cores of one single machine (Coro et al., 2017). After each computation, DataMiner saves - on the user's Workspace- all the information about the input and output data, and the experiment's parameters (computational provenance) using the Prov-O XML standard. In each D4Science VRE, DataMiner offers an online tool to integrate algorithms, which supports many programming languages (Coro et al., 2016). All these features make D4Science useful to develop OS-compliant applications, because WPS and provenance tracking allow repeating and reproducing a computation executed by another user. Also, the possibility to provide a process in multiple VREs focussing on different domains fosters its re-usability (Coro et al., 2017). In this paper, we will use the term "algorithm" to indicate processes running on DataMiner, and "method" to indicate the original processes or services integrated with DataMiner.

2.2 Annotations

NLPHub integrates a number of named entities recognizers (NERs) but also information extraction processes that recognize events, keywords, tokens, and sentences. Overall, we will use the term "annotation" to indicate all the information that NLPHub can extract from a text. The complete list of supported annotations, languages, and processes is reported in the supple-

mentary material, together with the list of all mentioned Web services' endpoints. The ontological classes used for NERs annotations come from the Standford CoreNLP software. Included non-standard annotations are "Misc" (miscellaneous concepts that cannot be associated with none of the other classes, e.g. "Bachelor of Science"), "Event" (nouns, verbs, or phrases referring to a phenomenon occurring at a certain time and/or space), and "Keyword" (a word or a phrase that is of great importance to understand the text content).

2.3 Integrated Text Mining Methods

NLPHub uses a common JSON format to represent the annotations of every integrated method. This format describes the input text, the NER processes, and the annotations for each NER:

```
"text": "input text",
  "NER_1": {
    "annotations":{
     "annotation_1":[
       {"indices": [i_1,i_2]},
       {"indices": [i_3,i_4]},
         ...,
```

We integrated services and methods with DataMiner through "wrapping algorithms" that transformed the original outputs into this format. We implemented a general workflow in each algorithm to execute the corresponding integrated method, which adopts the following steps: (i) receive an input text file and a list of entities to recognize (among those supported by the language), (ii) pre-process the text by deleting useless characters, (iii) encode the text with UTF-8 encoding, (iv) send the text via HTTP-Post to the corresponding service or execute the method on the local machine directly, if possible, and (v) return the annotation as an NLPHub-compliant JSON document. In the following, we list all the methods currently integrated with NLPHub with reference to Figure 1 for an architectural view.

CoreNLP. The Stanford CoreNLP software (Manning et al., 2014) is an open-source text processing toolkit that supports several languages (Stanford University, 2019). NLPHub integrates CoreNLP as a service instance running within D4Science with English, German, French, and Spanish language packages enabled. Also, the Tint (The Italian NLP Tool) extension for Italian (Aprosio and Moretti, 2016) was installed as a separate service. Overall, two distinct replicated and balanced virtual machines host these services on machines with 10 GB of RAM and 6 cores.

GATE Cloud. GATE Cloud is a cloud service that offers on-payment text analysis methods as-a-service (GATE Cloud, 2019a; Tablan et al., 2011). NLPHub integrates the GATE Cloud ANNIE NER for English, German, and French within a controlled VRE that accounts for users' requests load. This VRE ensures a fair usage of the services, whose access has been freely granted to D4Science in exchange for enabling OS-oriented features (SoBigData European Project, 2016).

OpenNLP. The Apache OpenNLP library is an open source text processing toolkit mostly based on machine learning models (Kottmann et al., 2011). An OpenNLP-based English NER is available as-a-service on GATE Cloud (GATE Cloud, 2019b) and is included among the free-to-use services granted to D4Science.

ItaliaNLP. ItaliaNLP is a free-to-use service - developed by the "Istituto di Linguistica Computazionale" (ILC-CNR) - hosting a NER method for Italian that combines rule-based and machine learning algorithms (ILC-CNR, 2019; Dell'Orletta et al., 2014).

NewsReader. NewsReader is an advanced events recognizer for 4 languages, developed by the NewsReader European project (Vossen et al., 2016). NewsReader is a formal inferencing system that identifies events by detecting their participants and time-space constraints. Two balanced virtual machines were installed in D4Science for the English and Italian NewsReader versions.

TagMe. TagMe is a service for identifying short phrases (*anchors*) in a text that can be linked to pertinent Wikipedia pages (Ferragina and Scaiella, 2010). TagMe supports 3 languages (English, Italian, and German) and D4Science already hosts its official instances. Since anchors are sequences of words having a recognized meaning within their context, NLPHub interprets them as keywords that can help contextualising and understanding the text.

Keywords NER. Keywords NER is an open-source statistical method that produces tags clouds of verbs and nouns (Coro, 2019a), which was also used by the H-Care award-winning human digital assistant (SpeechTEK 2010, 2019). Tag clouds are extracted through a statistical analysis of part-

of-speech (POS) tags (extracted with TreeTagger, (Schmid, 1995)) and the method can be applied to all the 23 TreeTagger supported languages. Keywords NER is executed directly on the DataMiner machines, and the nouns tags are interpreted as keywords for the NLPHub scopes, because - by construction - their sequence is useful to understand the topics treated by a text.

Language Identifier. NLPHub also provides a language identification process (Coro, 2019b), should language information not be specified as input. This process was developed in order to be fast, easily, and quickly extendible to new languages. The algorithm is based on an empirical behaviour of TreeTagger (common to many POS taggers): When TreeTagger is initialised on a certain language, but it processes a text written in another language, it tends to detect many nouns and unstemmed words than verbs and other lexical categories. Thus, the detected language is the one having the most balanced ratio of recognized and stemmed words with respect to other lexical categories. This algorithm is applicable to many languages supported by TreeTagger and can run on the DataMiner machines directly. An estimated accuracy of 95% on 100 sample text files covering the 5 NLPHub languages was convincing to use this algorithm as an auxiliary tool for the NLPHub users.

2.4 NLPHub

On top of the methods and services described so far, we implemented an alignment-merging algorithm (AMERGE) that orchestrates the computations and assembles their outputs. AMERGE receives a user-provided input text, along with the indication of the text language (optionally), and a set of annotations to be extracted (selected among those supported for that language). Then, it concurrently invokes - via WPS - the text processing algorithms that support the input request, and eventually collects the JSON documents coming from them. Finally, it aligns and merges the information to produce one overall sequence represented in JSON format. The issue of merging the heterogeneous connected services' outputs is solved through the use of the DataMiner wrapping algorithms. Another solved issue is the merge of the different intervals identified by several algorithms focusing on the same entities. These intervals may either overlap or be mutually inclusive, and the alignment algorithm manages all cases through algebraic evaluations, as reported in the following pseudo-code:

```
AMERGE Algorithm

For each annotation E:
 Collect all annotations detected
     by the algorithms (intervals
     with text start and end
    positions);
 Sort the intervals by their
    start position;
 For each segment s_i:
  If s_j is properly included in
     s_i, process the next s_j;
  If s_i does not intersect s_j,
     brake the loop;
  If s_i intersects s_j, create a
     new segment su_i as the union
     of the two segments →
     substitute su_i to s_i and
     restart the loop on s_j;
 Save s_i in the overall list of
    merged intervals S;
 Associate S to E;
Return all (E,S) pairs sets.
```

Since the AMERGE algorithm is a DataMiner algorithm, it is published as-a-service with a RESTful WPS interface. It represents one single access point to the services integrated with NLPHub. In order to invoke this service, a client should specify an authorization code in the HTTP request that identifies both the invoking user and the VRE (CNR, 2016). The available annotations and methods depend on the VRE. An additional service (NLPHub-Info) allows retrieving the list of supported entities for a VRE, given a user's authorization code. NLPHub is also endowed with a free-to-use Web interface (`nlp.d4science.org/hub/`), based on a public VRE, operating on top of the AMERGE process, which allows interacting with the system and retrieving the annotations in a graphical format.

3 Results

We assessed the NLPHub performance by using the I-CAB corpus as a reference (Magnini et al., 2006), which contains annotations of the following named-entities categories from 527 Italian newspapers: Person, Location, Organization,

	Person				Geopolitical				Location				Organization			
Algorithm	F-measure	Precision	Recall	Agreement	F-measure	Precision	Recall	Agreement	F-measure	Precision	Recall	Agreement	F-measure	Precision	Recall	Agreement
ItaliaNLP	79%	74%	84%	Excellent	77%	74%	80%	Good	59%	52%	69%	Good	58%	52%	66%	Good
CoreNLP-Tint	85%	78%	93%	Excellent	NA	NA	NA	NA	30%	18%	84%	Marginal	65%	53%	83%	Good
AMERGE	84%	74%	96%	Excellent	77%	74%	80%	Good	31%	19%	88%	Marginal	63%	49%	87%	Good
Keywords NER	20%	12%	56%	Marginal	14%	8%	66%	Marginal	6%	3%	58%	Marginal	22%	13%	66%	Marginal
TagMe	23%	18%	30%	Marginal	33%	22%	67%	Marginal	9%	5%	42%	Marginal	25%	19%	38%	Marginal
AMERGE - Keywords	20%	12%	69%	Marginal	18%	10%	91%	Marginal	6%	3%	74%	Marginal	22%	13%	79%	Marginal

Table 1: Performance assessment of the NLPHub algorithms with respect to the I-CAB corpus annotations.

Geopolitical entity. NLPHub was executed to annotate these same entities plus Keywords (Table 1). The involved algorithms were CoreNLP-Tint, ItaliaNLP, Keywords NER, and TagMe. According to the F-measure, CoreNLP-Tint was the best at recognizing Persons and Organizations, whereas ItaliaNLP - the only one supporting Geopolitical entities - had the highest performance on Locations and a moderately-high performance on Geopolitical entities. Overall, the connected methods showed high performance on specific entities, but there was not one method outperforming the others on all entities. AMERGE had lower but good F-measure and a generally high recall in all cases, which indicates that the connected algorithms include complementary and valuable intervals. The AMERGE-Keywords algorithm had a generally high recall (especially on Geopolitical entities), which means that the extracted keywords include also words from the annotated entities. The associated F-measures indicate that there is overlap with several entities. In turn, this indicates that AMERGE-Keywords could be a valuable source of information in the case of uncertainty about the entities that can be extracted from a text. As a further evaluation, we used Cohen's Kappa (Cohen, 1960) to explore the agreement between the algorithms and the I-CAB annotations. This measure required estimating the overall number of classifiable tokens, thus it is more realistic to refer to Fleiss' Kappa macro classifications rather than to the exact values (Fleiss, 1971). According to Fleiss' labels, all NERs generally have good agreement with I-CAB except for Locations, which are often reported as Geopolitical entities in I-CAB. This evaluation also highlights that AMERGE has good general agreement with manual annotations, and thus can be a valid choice when there is no prior knowledge about the algorithm to use for extracting a certain entity.

4 Conclusions

We have described NLPHub, a distributed system connecting and combining 30 text processing methods for 5 languages that adds Open Science-oriented features to these methods. The advantages of using NLPHub are several, starting from the fact that it provides one single access endpoint to several methods and spares installation and configuration time. Further, it proposes the AMERGE process as a valid option when the best performing algorithm for a certain entity extraction is not known *a priori*. Also, the AMERGE-Keywords annotations can be used when the entities to extract are not known. Indeed, these features would require more investigation, especially through multiple-language experiments, in order to define their full potential and limitations. Finally, NLPHub adds to the original methods features like WPS and Web interfaces, provenance management, results sharing, and access/usage policies control, which make the methods more compliant the with Open Science requirements.

The potential users of NLPHub are scholars who want to use NERs but also want to avoid software and hardware-related issues, or automatic agents that need to automatically extract and reuse knowledge from large quantities of texts. For example, NLPHub can be used in automatic ontology population and - since it also supports Events extraction - automatic narratives generation (Petasis et al., 2011; Metilli et al., 2019). Future extensions of NLPHub will involve other text mining methods (e.g. sentiment analysis, opinion mining, and morphological parsing), and additional NLP tasks like text-to-speech and speech processing as-a-service.

Supplementary Material

Supplementary material is available on D4Science at this permanent hyper-link.

References

[Adedugbe et al.2018] Oluwasegun Adedugbe, Elhadj Benkhelifa, and Russell Campion. 2018. A cloud-driven framework for a holistic approach to semantic annotation. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 128–134. IEEE.

[Amado et al.2018] Alexandra Amado, Paulo Cortez, Paulo Rita, and Sérgio Moro. 2018. Research trends on big data in marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*, 24(1):1–7.

[Andronico et al.2011] Giuseppe Andronico, Valeria Ardizzone, Roberto Barbera, Bruce Becker, Riccardo Bruno, Antonio Calanducci, Diego Carvalho, Leandro Ciuffo, Marco Fargetta, Emidio Giorgio, et al. 2011. e-infrastructures for e-science: a global view. *Journal of Grid Computing*, 9(2):155–184.

[Aprosio and Moretti2016] Alessio Palmero Aprosio and Giovanni Moretti. 2016. Italy goes to stanford: a collection of corenlp modules for italian. *arXiv preprint arXiv:1609.06204*.

[Ariadne2019] Ariadne. 2019. The AriadnePlus European Project. `https://ariadne-infrastructure.eu/`.

[Assante et al.2016] Massimiliano Assante, Leonardo Candela, Donatella Castelli, Gianpaolo Coro, Lucio Lelii, and Pasquale Pagano. 2016. Virtual research environments as-a-service by gcube. *PeerJ Preprints*, 4:e2511v1.

[Assante et al.2019] Massimiliano Assante, Leonardo Candela, Donatella Castelli, Roberto Cirillo, Gianpaolo Coro, Luca Frosini, Lucio Lelii, Francesco Mangiacrapa, Valentina Marioli, Pasquale Pagano, et al. 2019. The gcube system: Delivering virtual research environments as-a-service. *Future Generation Computer Systems*, 95:445–453.

[Bontcheva and Derczynski2016] Kalina Bontcheva and Leon Derczynski. 2016. Extracting information from social media with gate. In *Working with Text*, pages 133–158. Elsevier.

[Candela et al.2013] Leonardo Candela, Donatella Castelli, Gianpaolo Coro, Pasquale Pagano, and Fabio Sinibaldi. 2013. Species distribution modeling in the cloud. *Concurrency and Computation: Practice and Experience*.

[CNR2016] CNR. 2016. gcube wps thin clients. `https://wiki.gcube-system.org/gcube/How_to_Interact_with_the_DataMiner_by_client`.

[Cohen1960] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

[Coro et al.2015] Gianpaolo Coro, Leonardo Candela, Pasquale Pagano, Angela Italiano, and Loredana Liccardo. 2015. Parallelizing the execution of native data mining algorithms for computational biology. *Concurrency and Computation: Practice and Experience*, 27(17):4630–4644.

[Coro et al.2016] Gianpaolo Coro, Giancarlo Panichi, and Pasquale Pagano. 2016. A web application to publish r scripts as-a-service on a cloud computing platform. *Bollettino di Geofisica Teorica ed Applicata*, 57:51–53.

[Coro et al.2017] Gianpaolo Coro, Giancarlo Panichi, Paolo Scarponi, and Pasquale Pagano. 2017. Cloud computing in a distributed e-infrastructure using the web processing service standard. *Concurrency and Computation: Practice and Experience*, 29(18):e4219.

[Coro2019a] Gianpaolo Coro. 2019a. The Keywords Tag Cloud Algorithm. `https://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data-analysis/LatentSemanticAnalysis/`.

[Coro2019b] Gianpaolo Coro. 2019b. The Language Identifier Algorithm. hyper-link.

[Dell'Orletta et al.2014] Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. T2kˆ 2: a system for automatically extracting and organizing knowledge from texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

[EU Commission2016] EU Commission. 2016. Open science (open access). `https://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access`.

[Ferragina and Scaiella2010] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.

[Fleiss1971] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

[Gandomi and Haider2015] Amir Gandomi and Murtaza Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.

[GATE Cloud2019a] GATE Cloud. 2019a. GATE Cloud: Text Analytics in the Cloud. `https://cloud.gate.ac.uk/`.

[GATE Cloud2019b] GATE Cloud. 2019b. OpenNLP English Pipeline. https://cloud.gate.ac.uk/shopfront/displayItem/opennlp-english-pipeline.

[Hey et al.2009] Tony Hey, Stewart Tansley, Kristin M Tolle, et al. 2009. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA.

[ILC-CNR2019] ILC-CNR. 2019. The ItaliaNLP REST Service. http://api.italianlp.it/docs/.

[Kottmann et al.2011] J Kottmann, B Margulies, G Ingersoll, I Drost, J Kosin, J Baldridge, T Goetz, T Morton, W Silva, A Autayeu, et al. 2011. Apache OpenNLP. *www.opennlp.apache.org*.

[Lebo et al.2013] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. 2013. Prov-o: The prov ontology. *W3C Recommendation*, 30.

[Linthicum2017] David S Linthicum. 2017. Cloud computing changes data integration forever: What's needed right now. *IEEE Cloud Computing*, 4(3):50–53.

[Magnini et al.2006] Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi, and Rachele Sprugnoli. 2006. I-cab: the italian content annotation bank. In *LREC*, pages 963–968. Citeseer.

[Manning et al.2014] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

[Metilli et al.2019] Daniele Metilli, Valentina Bartalesi, and Carlo Meghini. 2019. Steps towards a system to extract. In *Proceedings of the Text2Story 2019 Workshop*, page na. Springer.

[OpenAire2019] OpenAire. 2019. European project supporting Open Access. https://www.openaire.eu/.

[OpenMinTeD2019] OpenMinTeD. 2019. Open Mining INfrastructure for TExt and Data. https://cordis.europa.eu/project/rcn/194923/factsheet/en.

[Parthenos2019] Parthenos. 2019. The Parthenos European Project. http://www.parthenos-project.eu/.

[Petasis et al.2011] Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, and Elias Zavitsanos. 2011. Ontology population and enrichment: State of the art. In *Knowledge-driven multimedia information extraction and ontology evolution*, pages 134–166. Springer-Verlag.

[Pollock and Williams2010] Neil Pollock and Robin Williams. 2010. E-infrastructures: How do we know and understand them? strategic ethnography and the biography of artefacts. *Computer Supported Cooperative Work (CSCW)*, 19(6):521–556.

[Schmid1995] Helmut Schmid. 1995. Treetagger - a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

[Schut and Whiteside2007] Peter Schut and A Whiteside. 2007. OpenGIS Web Processing Service. OGC project document http://www.opengeospatial.org/standards/wps.

[SoBigData European Project2016] SoBigData European Project. 2016. Deliverable D2.7 - IP principles and business models. http://project.sobigdata.eu/material.

[SoBigData2019] SoBigData. 2019. The SoBigData European Project. http://sobigdata.eu/index.

[SpeechTEK 20102019] SpeechTEK 2010. 2019. SpeechTEK 2010 - H-Care Avatar wins People's Choice Award. http://web.archive.org/web/20160919100019/http://www.speechtek.com/europe2010/avatar/.

[Stanford University2019] Stanford University. 2019. Stanford CoreNLP - Human Languages Supported. https://stanfordnlp.github.io/CoreNLP/.

[Tablan et al.2011] Valentin Tablan, Ian Roberts, Hamish Cunningham, and Kalina Bontcheva. 2011. GATE Cloud.net: Cloud Infrastructure for Large-Scale, Open-Source Text Processing. In *UK e-Science All hands Meeting*.

[Vossen et al.2016] Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, et al. 2016. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85.

[Wei et al.2016] Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. 2016. Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics*, 32(12):1907–1910.