Reusable Research?
A Case Study in Named Entity Recognition

Marieke van Erp[1] and Lourens van der Meij[2]

The Network Institute
[1] Computational Lexicology and Terminology Lab
[2] Department of Computer Science
VU University Amsterdam, the Netherlands
{marieke.van.erp,lourens.vander.meij}@vu.nl

Computational Lexicology and Terminology Lab
Technical Report CLTL 2013-01

April 2013

Abstract. Named Entity Recognition (NER) is a crucial piece of knowledge acquisition infrastructure for the Semantic Web. Thus, being able to build-upon and reuse state-of-the art approaches in this domain is essential for continued progress. In this paper, we present results in attempting to reproduce the algorithms presented by Freire et al. in ESWC 2012 [1]. Based on this result, we come to the conclusion that even if experiments are described in detail it is still very difficult to reproduce the experiments and pinpoint the particular difficulties in this use case. Finally, we evaluate our attempts as well as those of [1] on the named entity recognition task.

1 Introduction

As a research community, we build upon each other's work. We collaborate with each other, describe our systems, and make available our data and software for reuse. This sharing of research data, methods, and results is imperative for progress. However, even though as a community we strive to make our results reusable [2,3], we are still quite far from a world in which you can pick up a paper, rerun the experiments described in that paper, and build upon these results within a reasonable timeframe.

In this report, we describe a use case in which we attempt to reproduce an approach and experiments described at last year's ESWC conference. We chose to try to reproduce the experiments described in [1] as they describe an approach for domain adaptation for named entity recognition (NER) for the cultural heritage domain. NER is the task of identifying and classifying named entities in text, one of the most crucial tasks in annotating data. The authors could not make their system available, but they have made their annotated data available and their paper describes their system and experiments in great detail.

Our experiments show that, even with detailed system descriptions, available data and some help from the authors, it is still very difficult to reproduce someone else's research experiments. In this report, we analyse the different stages of reimplementing someone else's approach, we attempt to explain the reason for our different results.

We also show that despite the fact that we could not entirely reproduce the results described in [1], taking on a reimplementation task can still provide useful insights into the NER task, as well as some improvements to the experiments and result analysis.

The remainder of this report is organised as follows. In Section 2, we detail the motivation for replicating these experiments and why we chose this particular domain adaptation. In Section 3, the approach and experiments we attempted to replicate are described, followed by our replication experiment in Section 4. Even though we did not manage to replicate [1] results, we could still carry out some analyses and extra experiments. These are described in Section 5. We conclude with conclusions in Section 6.

2 Motivation

Reproducibility of research results is not a new topic in the academic debate. The importance of reproducibility is taught to both graduate and undergraduate students, and most academics will have also encountered it in their discussions with colleagues. Analyses of the issues preventing reproducibility as well as recommendations for improving it are myriad in the literature (cf. [2,4,5,3]).

There are also community-driven initiatives and incentives, such as the reproducibility track in the VLDB conferences[1] and enouragement of accompanying papers with software and/or data various conferences such as ESWC[2] and ACL[3]. Furthermore, publishers are working on publication models that make publication of research data and/or software with articles easier[4,5]. Despite these efforts, most research is still very difficult to reproduce barring quick uptake.

Many domains would benefit from being able to tap into research results quicker. One such domain is the cultural heritage domain, which harbours vast amounts of data that are valuable to the data owners (cultural heritage institutions), humanities researchers as well as the general public. In recent years, many cultural heritage institutions have started to digitise their data for preservation purposes, as well as improved data access internally and to the general public [6].

As research in the humanities often revolves around particular persons in time, NER is a good start for any tool designed to help such researchers struc-

[1] http://www.vldb.org/conference.html

[2] http://2013.eswc-conferences.org/

[3] http://http://acl2013.org/

[4] http://thoughtsonpublishing.wordpress.com/2012/01/12/scholarly-enrichments/

[5] http://figshare.com/

ture their resources. In many instances in the museum domain, events, and their associated persons, locations and times, also make up a large part of the interesting clusters to be found in their collections. It is with questions from both humanities researchers and museum professionals looking to gain more structured access to their data that we set out on exploring the options for NER in the cultural heritage domain.

In the next section, we summarise the approach and results presented in [1], before we present our reproducibility experiments in Section 4.

3 Target of Reproducibility

The approach presented in [1] aims to resolve the domain adaptation problem for named entity recognition by using a statistical method with complex features that are based on the domain-specific data. They argue that by computing frequency statistics over a large number of person and organisation names and using these as features, one can reliably train a classifier with a small portion of training data. Their domain data set consists of 120 records describing objects from cultural heritage, provided by European libraries, museums and archives. This data set is compiled in the framework of the Europeana project[6]. Each record describes an object from one of the heritage institutions participating in Europeana. As can be seen in Figure 1, this information ranges from the type and dimensions of the object to a description of what is depicted on the object or what it represents.

One of the distinguishing features of the work of [1] in NER is that they treat the fact that the Europeana records are a mix of structured and unstructured text as a feature rather than a bug. They employ knowledge about the structure of the records as a feature in the system. For the evaluation [1], annotated the seven different parts of the Europeana records (data elements) with Location, Organisation and Person information and made this dataset available for reuse.

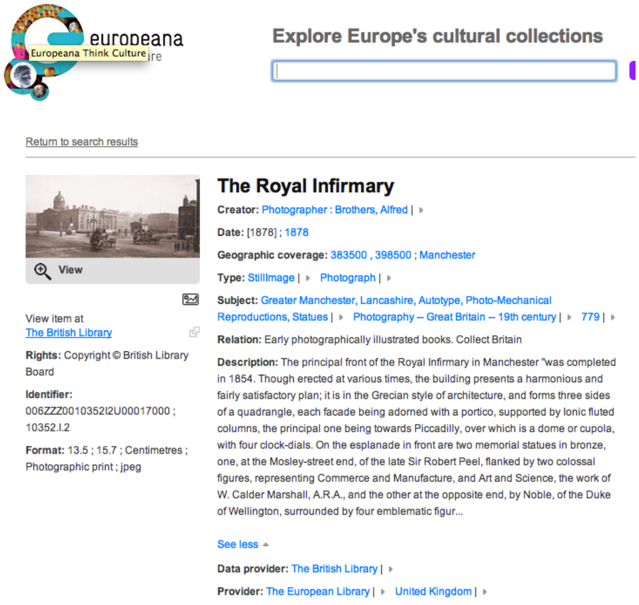
The core of the [1] approach is the following. First, a set of complex features is generated for each token based on domain knowledge about person and organisation names from VIAF[7] and about location names from GeoNames[8]. In these features, knowledge is encoded about for example how often a token occurs as a first name or a location name in the external resource. WordNet [7][9] is used to capture information about the possible past-of-speech tags a token might have, for example, if a token occurs as a noun in WordNet then the binary feature for noun is set to 1. A window of three preceding and one following token is also included, as well as statistics about the proportion of capitalised tokens in the data set. Standard features such as whether a token is capitalised, allcaps, is at the beginning or end of a data element and its length are also included. The last

[6] `http://www.europeana.eu/`

[7] `http://viaf.org/`, last queried 5 July 2012

[8] `http://www.geonames.org`, last queried 14 December 2012

[9] `http://wordnet.princeton.edu/`



The Royal Infirmary

Creator: Photographer : Brothers, Alfred |

Date: [1878] ; 1878

Geographic coverage: 383500 , 398500 ; Manchester

Type: StillImage | Photograph |

Subject: Greater Manchester, Lancashire, Autotype, Photo-Mechanical Reproductions, Statues | Photography -- Great Britain -- 19th century | 779 |

Relation: Early photographically illustrated books. Collect Britain

Description: The principal front of the Royal Infirmary in Manchester "was completed in 1854. Though erected at various times, the building presents a harmonious and fairly satisfactory plan; it is in the Grecian style of architecture, and forms three sides of a quadrangle, each facade being adorned with a portico, supported by Ionic fluted columns, the principal one being towards Piccadilly, over which is a dome or cupola, with four clock-dials. On the esplanade in front are two memorial statues in bronze, one, at the Mosley-street end, of the late Sir Robert Peel, flanked by two colossal figures, representing Commerce and Manufacture, and Art and Science, the work of W. Calder Marshall, A.R.A., and the other at the opposite end, by Noble, of the Duke of Wellington, surrounded by four emblematic figur...

Data provider: The British Library |

Provider: The European Library | United Kingdom |

View item at
The British Library

Rights: Copyright © British Library Board

Identifier: 006ZZZ0010352I2U00017000 ; 10352.I.2

Format: 13.5 ; 15.7 ; Centimetres ; Photographic print ; jpeg

Fig. 1. Example of a Europeana Data Record

feature encodes in which data element the token occurs, e.g., whether it is in the record's title or in the description element.

Second, a conditional random field classifier (CRF) [8] is trained and run on the data with the generated features in a 10-fold cross-validation experiment. The CRF implementation that was used was Mallet - Machine Learning for Language Toolkit [9][10] that used the three previous states in the sequence as well as a the label likelihood and a Gaussian prior on its parameters.

Finally, they evaluated their approach using the CoNLL NER shared task evaluation methodology [10] that only rewards a decision made by the classifier if it correctly classifies the entire entity, thus not awarding points for only recognition of part of the entity. Results are reported using precision, recall and $F_{\beta=1}$. They compare their approach against an off-the-shelf approach, the Stanford named entity recogniser[11][11] trained on the CoNLL English NER shared task data [10].

A summary of the results Freire et al. obtained as well as their Stanford baseline results is given in Table 1. In this table, we have only listed the results of their system that had a confidence of .9.

In the next section, we explain how we tried to reproduce the results in [1] in order to be able to utilise their approach for annotating cultural heritage records with named entities.

[10] http://mallet.cs.umass.edu
[11] http://nlp.stanford.edu/software/CRF-NER.shtml

Table 1. Results of Freire et al. 2012 as summarised from Figures 1 and 2 in [1]

	Freire et al.			Freire et al. Stanford baseline		
	precision	recall	$F_{\beta=1}$	precision	recall	$F_{\beta=1}$
Persons	.92	.55	.69	.69	.22	.33
Organisations	.90	.57	.70	.42	.14	.21
Locations	.91	.56	.69	.81	.08	.15
Overall	.91	.55	.69	.71	.14	.23

4 Replication Experiments

In this section, we will detail the process of reimplementing the approach described in [1], as well as our evaluations and explanations of the discrepancy in results. Our code, data and experimental settings can be found at: `https://github.com/MvanErp/NER`.

Reimplementing the approach Unfortunately, the code for the approach from [1] was not available, but as the paper provided detailed explanations of how their complex features were computed, as well as their annotated data set, it seemed relatively straightforward to re-implement their approach.
Complex features person and organisation names For the personFirstName, personSurname, personNoCapitalsName and organisationName, statistics are computed from the the VIAF authority file[7] to create the features. These describe for example how often a token occurs as a first name of a person. As VIAF is structured around 1 record per entity, there may be several alternative names for an entity. We split the person names between first name and last name and removed other formatting such as parentheses and dates from the names. We gathered statistics on also all the alternative names for entities, but initials were not included.
Complex features GeoNames Statistics for locationName are computed using GeoNames[8], using the perl Geo::GeoNames module[12].
WordNet features Binary values indicating whether a token may be a properNoun, noun (posNoun), verb (posVerb), adjective (posAdjective) or adverb (posAdverb) are found using WordNet [7]. We used WordNet 3.0 which we queried with the perl WordNet::QueryData module[13]. Through this setup it was not possible to query WordNet for part of speech tags for prepositions and proper nouns as [1] describes. Therefore we could not include these features.
Other features The other features included are fairly generic such as whether a token is capitalised or not, whether it is at the start or end of a data element, its length etc. For a full description of the features, the reader is referred to [1].

We received some extra information from the authors about the features, specifically about for which tokens around the focus token certain features such

[12] `http://search.cpan.org/~perhenrik/Geo-GeoNames/lib/Geo/GeoNames.pm`, version 0.11

[13] `http://search.cpan.org/dist/WordNet-QueryData/QueryData.pm`, version 1.49

Table 2. Description of which features for and around focus token and window size

Feature name	Previous	Following	Feature name	Previous	Following
token	3	3	posVerb	2	1
personFirstName	1	1	posAdjective	2	1
personSurname	1	1	posAdverb	2	1
personNoCapitalsName	1	1	isCapitalised	2	2
organisationName	1	1	isUppercased	2	2
locationName	1	1	tokenLength	2	2
capitalisedFrequency	1	1	startOfElement	0	0
properNoun	2	1	endOfElement	0	0
posNoun	2	1	dataElement	0	0

as capitalisation are also computed. This resulted in a feature vector containing the token to be classified, 57 features providing extra information, plus the named entity class. In Table 2, we show the window size that was used for each feature[14]. Some features are only computed for the previous and following token, whereas others are created for the three preceding and three following tokens.

Running the experiments As Freire et al., we use the Mallet [9][15] machine learning toolkit to train and test the models. We used the command line interface in Mallet version 2.0.7.

4.1 Replicating the Stanford baseline

The Stanford Named Entity Recogniser[16][11] is a statistical named entity recogniser that also uses a CRF classifier. On domains for which it has been trained (i.e., the training data is of the same genre and similar topics to the test data) its performance is state of the art, with an overall F_1 of 86.86. When applied to different domains, performance drops significantly. Our experiments were carried out with the Stanford Named Entity Recognizer 1.2.5, released on 2012-05-22.

4.2 Replication results

In Table 3, the results of our reproduction experiments are presented. After trying several variations with the feature sets, this was the closest we could get to their experiments. What is more interesting however, is that even replicating the Stanford results yielded different results. We could not test whether these differences are significant, as we did not have access to the output of the system described in [1].

[14] The tokens preceding and following the token to be classified are concatenated and represented as a single feature

[15] `http://mallet.cs.umass.edu`

[16] `http://nlp.stanford.edu/software/CRF-NER.shtml`

Table 3. Precision, recall and $F_{\beta=1}$ measures for our replication of the Freire et al. 2012 approach and Stanford parser trained on CoNLL 2003 data

	Freire et al. Replication			Stanford CoNLL 2003		
	Precision	Recall	$F_{\beta=1}$	Precision	Recall	$F_{\beta=1}$
LOC (388)	77.80%	39.18%	52.05	65.17%	55.93%	60.19
ORG (157)	65.75%	30.57%	41.74	4.33%	15.92%	6.81
PER (614)	73.33%	37.62%	49.73	59.88%	50.33%	54.69
Overall (1,159)	73.33%	37.19%	49.45	34.42%	47.54%	39.93

4.3 Discrepancy Explained

Implementing their features and fine-tuning the approach took considerable time and effort ($>$ 1.5 months) and did not yield a perfect replication of the results. We have defined four possible causes for our inability to reproduce the experiments of [1] exactly.

Data Conversion The original data was in .xml format, whereas for the evaluation, the text in each of the .xml elements was to be tokenised and the named entities converted to the iob1 format [12]. There is a significant difference between the output of different tokenisation algorithms. We used a very simple perl module to tokenise the text[17]. There are more linguistically informed tokenisers available for English, but as the text in some of the data elements was not grammatically well-formed we opted for an approach that is less likely to suffer from ungrammaticality.

Feature Generation In order to extract the person and organisation name features from VIAF, it was necessary to first identify persons and organisations in VIAF and to split the names in for example first and last names. Design choices made in the module we developed most likely will have resulted in slightly different frequency values, for example on whether to include names in non-Western characters or not, and variation may have also occurred in splitting the different parts of names.

Feature Formatting Although description of their features and classifier settings is fairly elaborate, there were still settings that we had to guess at such as the number of decimals places to use to compute the features (we chose 3 in the end).

Machine Learning Settings We chose to use the Mallet machine learner through the command line interface. More sophisticated fine-tuning is possible by digging into the java code. As we could only guess at the design choices made in [1], we decided to use the standard implementation.

Reproducing the Stanford Baseline In [1] the baseline against which their system is tested is the Stanford NER system version 1.2.5 trained on the CoNLL 2003 dataset. There are three different models provided with the Stanford NER system that are trained on the CoNLL 2003 data. We have run all three models,

[17] `http://search.cpan.org/~andrefs/Lingua-EN-Tokenizer-Offsets-0.01_03/lib/Lingua/EN/Tokenizer/Offsets.pm` version 0.01_03

but our results are closest to the results reported in [1] with the conll.distsim.iob2.crf.ser.gz model. In order to run this model, we first needed to provide part-of-speech tags for the data, which we did by running the Stanford POS tagger that was released on 2012-11-11 with the `english-bidirectional-distsim.tagger`. Differences in our results for the Stanford baseline may stem from our tokenisation, part-of-speech tagging, or chosen NER model.

Despite the fact that we could not reproduce the system in [1], we could perform additional experiments and analyses to gain more insights into domain adaptation for NER. In the next section, our experiments and findings are detailed.

5 The Benefits of Reproduction

When digging into the approach and experiments carried out by [1], we set up various other experiments to gain greater insights into the workings of their approach and the peculiarities of the dataset used. In this section we detail our results in these explorations and what this could mean for the NER domain adaptation task. We first describe our experiments, followed by results and error analysis. We conclude this section by a discussion and suggestions for further experiments.

5.1 The Europeana Dataset

As mentioned in Section 3, the authors in [1] employ the fact that the data in the Europeana dataset is not heterogeneous. However, it may be the case that the structure of the different data elements may influence the results negatively. In order to investigate this, we first analysed the distribution of named entities over the different data elements, these statistics are given in Table 4[18].

As the statistics show, the contents of the data elements are quite different. As the descriptions usually contain running text, they make up for nearly half of the total dataset. However, the proportion of named entities in this data element is much lower than in other data elements (9.09%), with the publisher data element being the most different in this respect as it named entities make up more than half of the data in this data element (53.86%).

The distribution of the types of named entities also varies over the different data elements; person names make up the majority of the named entities in the titles, creator, and description elements, whereas locations make up the majority class of entities in the subject and publisher classes. Organisations make up the smallest named entity class in this dataset with only 157 in total, fewer than the person class in the description elements alone. They do make up a large part of the named entities in the publisher element.

[18] [1] also list a coverage element, but as it only occurs once in the dataset and does not contain any NEs it is not listed here.

Table 4. Statistics of the Europeana named entity evaluation data set. Per data element, the number of tokens, unique tokens, named entities (phrases), named entity tokens and non-named entity tokens is given. For the latter two, also the proportions are given. Also, the number of named entities (phrases) of the different named entity types are presented.

Element	Tokens	Unique	NEs	NEtoken	No NE	LOC	ORG	PER
Title	2,640	510	268	497(18.83%)	2,143 (81.17%)	87	28	153
Creator	1,065	177	197	371(34.84%)	694 (65.16%)	16	25	156
Subject	1,378	257	215	268(19.45%)	1,110 (80.55%)	129	16	70
Description	6,286	983	313	571(9.09%)	5,715 (90.91%)	75	38	200
ToC	621	319	44	70(11.28%)	551(88.72%)	30	2	12
Publisher	505	115	111	272(53.86%)	233 (46.14%)	52	42	17
Total	12,510	3,701	1,159	2,049(16.38%)	10,461 (83.62%)	388	157	614

Table 5. Precision, recall and $F_{\beta=1}$ measures for overall text

	Stanford Europeana Data no POS			Stanford Europeana Data + POS		
	Precision	Recall	$F_{\beta=1}$	Precision	Recall	$F_{\beta=1}$
LOC (388)	81.58%	63.92%	71.68	82.52%	65.72%	73.17
ORG (157)	74.42%	40.76%	52.67	64.13%	37.58%	47.39
PER (614)	75.09%	69.22%	72.03	76.88%	68.24%	72.30
Overall (1,159)	77.09%	63.59%	69.69	77.48%	63.24%	69.64

5.2 Experiments and Results

In [1] the baseline against which is tested is a system that is trained on newswire. However, to gain insight into the influence of the complex features that are defined, we devised an experiment in which we retrained the Stanford Named Entity recogniser on the Europeana data. As [1], we performed a 10-fold cross validation experiment. In this experiment, we only used the tokens and the named entity class. To train the models, the same settings were used as for the CoNLL/MUC 3-class model[19]. As in the CoNLL 2003 experiments, POS tags were also supplied, we also performed a set of 10-fold cross-validation experiments in which our dataset contained tokens and POS tags[20]. The results of these experiments are presented in 5. However, these results are slightly worse than those without POS tags, so in our analyses we focused on the Stanford experiments without part-of-speech tags.

As Table 4 showed, there is a fair difference in the length and distributions of named entities in the different data elements. Although this dataset is fairly small, and thus it is difficult to draw general conclusions from it, it can be assumed that the class distributions influence named entity results. Therefore, we also report results for each data element separately in Tables 6-11.

[19] Included in the Stanford release in the `english.all.3class.distsim.prop` file.

[20] For this we also used the Stanford POS tagger, release 2012-11-11

Table 6. Precision, recall and $F_{\beta=1}$ measures for title data element

	Freire et al. Replication			Stanford Europeana Data		
	Precision	Recall	$F_{\beta=1}$	Precision	Recall	$F_{\beta=1}$
LOC (87)	69.57%	32.00%	43.84	90.32%	56.00%	69.14
ORG (28)	60.00%	30.00%	40.00	50.00%	20.00%	28.57
PER (153)	53.85%	32.56%	40.58	72.97%	62.79%	67.50
Overall (268)	61.11%	32.04%	42.04	79.17%	55.34%	65.14

Table 7. Precision, recall and $F_{\beta=1}$ measures for creator data element

	Freire et al. Replication			Stanford Europeana Data		
	Precision	Recall	$F_{\beta=1}$	Precision	Recall	$F_{\beta=1}$
LOC (16)	100.00%	28.57%	44.44	100.00%	28.57%	44.44
ORG (25)	0%	0%	0	0%	0%	0
PER (156)	89.47%	80.95%	85.00	75.00%	71.43%	73.17
Overall (197)	90.48%	57.58%	70.37	77.27%	51.52%	61.82

Some data elements, such as the creator and publisher elements are quite homogeneous, which makes it easier to process for natural language processing tools. The creator class for example, almost always contains a person name (e.g.,*Waldron, Laurence* or *Albert Tezla*) , and the publisher class is usually made up of an organisation and a location (e.g., *Longmans (London)*). This is reflected in the high performance scores for locations and organisations in Tables 7 and 11. Another class that performs well, but is a slightly less homogeneous is the subject class (see Table 8), this class contains values such as *Ireland – School children* and *Kennedy, John F. (John Fitzgerald), 1917-1963*. Although there is more variation in the types of values that this data element holds, the person names and locations do follow fairly standardised formatting.

The description data element is most similar to the type of text one finds in the majority of NER research as the text contained in it consists of full grammatical sentences. This element makes up about half of the entire data set and displays much linguistic variation, as descriptions may describe the life of the artist who created the object that is described, or finer details of the object such as what it depicts or its dimensions. It seems that particularly for this class, the more standard approach of only training on tokens is more successful. This is probably due to the fact that the advanced features make the feature set so complex and sparse that the classifier has trouble generalising over them. Perhaps a larger data set would mitigate this problem, but this remains to be tested.

One of the problems with the data set is that it is rather small, and thus it is difficult to generalise from the presented results. This holds in particular for the table of contents data element which only contains two organisation entities, making the sample size too small for the classifiers to do create a well-informed model.

Table 8. Precision, recall and $F_{\beta=1}$ measures for subject data element

	Freire et al. Replication			Stanford Europeana Data		
	Precision	Recall	$F_{\beta=1}$	Precision	Recall	$F_{\beta=1}$
LOC (129)	85.00%	44.74%	58.62	93.10%	71.05%	80.60
ORG (16)	60.00%	37.50%	46.15	60.00%	37.50%	46.15
PER (70)	83.33%	45.45%	58.82	54.55%	54.55%	54.55
Overall (268)	80.65%	43.86%	56.82	80.00%	63.16%	70.59

Table 9. Precision, recall and $F_{\beta=1}$ measures for description data element

	Freire et al. Replication			Stanford Europeana Data		
	Precision	Recall	$F_{\beta=1}$	Precision	Recall	$F_{\beta=1}$
LOC (75)	64.29%	32.73%	43.37	57.14%	58.18%	57.66
ORG (38)	30.00%	11.54%	16.67	85.71%	23.08%	36.36
PER (200)	76.19%	44.04%	55.81	68.75 %	70.64%	69.68
Overall (313)	68.32%	36.32%	47.42	65.71%	60.53%	63.01

5.3 Error Analysis

In our error analysis, we first start by looking at the prediction distribution. In Table 12, we have shown for each class in the gold standard the number of times the approaches predict that class or confuse it with another class. In Table 12 the class confusions are shown on the token level. As Table 12 shows, there are no classes that stand out in the sense that they often get confused. In most cases where the approaches go wrong, a named entity is predicted when there is none (O) or no named entity is predicted.

When we look at the cases where the approaches miss a named entity, i.e., where it erroneously predicts O instead of a named entity we find the following causes:

Ungrammatical phrases: In particular in the shorter data elements quite a few shorter phrases and parentheses are used. Both approaches seem to have difficulties with recognising phrases within parentheses or following “;”, “]”, “–”. This is probably due to the variation that can be contained within such markup and the data elements being too short to have other evidence to base its decision on.

Foreign phrases: The majority of the corpus is English, but sometimes foreign phrases such as French or Gaelic are encountered (in particular in the subject data element), and phrases such as *Eibhlín Nic Diarmada* or *Nouvelle-Zélande* are not recognised as named entities

Ambiguous phrases: if a term also occurs frequently as non-named entity, for example ‘South’ it may get missed when it is part of a named entity (e.g., *South Carolina*).

Long phrases: In particular for the location named entities, it holds that the majority of the named entities consists of single token entities. Longer entities, in particular those containing a phrase which on its own can also constitute a

Table 10. Precision, recall and F$_{\beta=1}$ measures for table of contents data element

	Freire et al. Replication			Stanford Europeana Data		
	Precision	Recall	F$_{\beta=1}$	Precision	Recall	F$_{\beta=1}$
LOC (30)	85.71%	50.00%	63.16	85.00%	70.83%	77.27
ORG (2)	100.00%	53.85%	70.00	88.89%	61.54%	72.73
PER (12)	65.00%	59.09%	61.90	73.08%	86.36%	79.17
Overall (44)	78.05%	54.24%	64.00	80.00%	74.58%	77.19

Table 11. Precision, recall and F$_{\beta=1}$ measures for publisher data element

	Freire et al. Replication			Stanford Europeana Data		
	Precision	Recall	F$_{\beta=1}$	Precision	Recall	F$_{\beta=1}$
LOC (52)	66.67%	18.18%	28.57	85.71%	54.55%	66.67
ORG (42)	75.00%	25.00%	37.50	50.00%	16.67%	25.00
PER (17)	14.29%	25.00%	18.18	20.00%	50.00 %	28.57
Overall (111)	42.86%	22.22%	29.27	47.62%	37.04%	41.67

named entity such as *Tower of London* are difficult to recognise completely. It must be said that the gold standard does not seem entirely consistent either as *Tower of London* is marked up as a location entity, but in *Lakes of Killarney* only *Killarney* is marked up. This may make it more difficult for the classifier to recognise such entities.

The counter examples, the cases in which the approaches predict an entity where there is none are largely explained by the following (similar) causes:

Terms derived from named entities: Some non-named entities are very similar to named entities or derived from them, such as *Icelandic*. In the CoNLL task, such phrases would be tagged with the tag MISC, but in our setup we do not have such a tag.

Foreign Phrases: Foreign phrases, in particular if they are capitalised may be mistaken for named entities, such as "Chasseur" in *Ténor (Chasseur) McLoughlin*

Other entities: Some phrases that we found refer to some kind of entity that is not contained in our entity classes, such as *Age of Realism*, which denotes a literary period, or *Death of Buta* which could be considered an event. These phrases 'look' quite similar to named entities such as "Tower of London" in the usage of capitalisation and prepositions which confuses the classifiers.

Upon inspection of the results, we also found some errors in the gold standard annotations. As was mentioned in the confusion with non-named entities, sometimes longer phrases containing locations are marked up, whereas sometimes only part is marked up. In some cases, we found a named entity markup missing such as in *Margaret Stokes (Carrig Brear, Howth)* in which only "Margaret Stokes" is marked up, and not Carrig Brear, Howth, which is a location in Ireland. As the error rate of human annotators lies between 2.5 and 3%[13], it is not surprising that there are some inconsistencies in the data, but it does

Table 12. Class prediction distribution for Freire et al. and Stanford approaches trained and tested on the Europeana data

	Freire et al. Replication				Stanford Retrained			
	LOC	ORG	PER	O	LOC	ORG	PER	O
LOC	**188**	7	18	279	**300**	3	36	153
ORG	9	**218**	21	232	14	**248**	28	190
PER	8	6	**437**	626	13	10	**799**	255
O	22	21	151	**10,270**	32	46	119	**10,264**

affect the approaches somewhat as they are fed erroneous examples and there is not that much data to go by in the first place.

5.4 Discussion

The size of the dataset is fairly small (12,510 tokens in total, and a little over 1,000 named entities). In particular when one looks at the individual data elements, the number of named entities rapidly becomes too small to draw any general conclusions from, with perhaps the exception of the description class. However, we do feel it is necessary to analyse the performance of the classifiers on the different elements individually as there are significant differences in the distributions of named entities over the different data elements.

The most interesting thing to note is that the advanced features are still often surpassed by a simple setup in which a model only trained on tokens. Even when comparing the results to the original results reported by Freire et al. the classifier trained on just the tokens from the Europeana data outperforms the Freire et al. results for recall and $F_{\beta=1}$ for locations, persons and overall. Indeed higher precision results are not obtained, although this may be mitigated by increasing the amount of training data. As in [1] scores are not reported for individual data elements we could not compare our results to theirs on this level, but our experiments indicate that models trained on only tokens are quite robust and can cope quite well with smaller amounts of training data. In particular for the 'smaller' data elements (i.e., those making up a smaller proportion of the dataset) the Stanford classifier outperforms the [1] approach significantly, probably because it is unaware of the fact that this data belongs to a different element.

More annotated training data will make it possible to draw more general conclusions about the influence of advanced features on the performance of NER approaches as well as training separate classifiers for the different data elements (e.g., one classifier for titles, one for descriptions etc.).

6 Conclusion

We have presented a use case in reproducibility of NER results for the cultural heritage domain. Our experiments and analyses show that it even if detailed system descriptions and experimental data are available it is still very difficult to reproduce the experiments. Fortunately, the research community is well aware of this fact and more tools are being made available to make it easier to share research data, software and results.

Acknowledgments
NWO in the CATCH programme, grant 640.004.801. The authors thank their colleagues for her helpful suggestions.

References

1. Freire, N., Borbinha, J., Calado, P.: An approach for named entity recognition in poorly structured data. In: Proceedings of ESWC 2012. (2012)
2. Faniel, I.M., Zimmerman, A.: Beyond the data deluge: A research agenda for large-scale data sharing and reuse. The International Journal of Digital Curation **6**(1) (2011) 58–69
3. Bechhofer, S., Buchan, I., DeRoure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., Goble, C.: Why linked data is not enough for scientists. Future Generation Computer Systems **29**(2) (Feb 2013) 599–611
4. Pederson, T.: Empiricism is not a matter of faith. ACL **34**(3) (2008) 465–470
5. Mesirov, J.P.: Accessible reproducible research. Science **327**(5964) (Jan 2010) 415–416
6. Karp, C.: Digital heritage in digital museums. Museum International **56**(1-2) (2004) 45–51
7. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. The MIT Press (1998)
8. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williamstown, MA, USA, Morgan Kaufmann (June 28 - July 1 2001) 282–289
9. McCallum, A.K.: MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu` (2002)
10. Tjong Kim Sang, E.F., Meulder, F.D.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL-2003, Edmonton, Canada (2003) 142–147
11. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005). (2005) 363–370
12. Tjong Kim Sang, E.F., Veenstra, J.: Representing text chunks. In: Proceedings of the 9th Conference of the european Chapter of the Association for Computational Linguistics (EACL 1999), Bergen, Norway (June 1999) 173–179
13. Marsh, E., Perzanowski, D.: MUC-7 evaluation of ie technology: Overview of results. In: Proceedings of MUC-7. (1998)