ACERPI: An approach for ordinances collection, information extraction and entity resolution

Christian Schmitz[1] **, Serigne K. Mbaye**[2]**, Edimar Manica**[2]**, Renata Galante**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

[2]Instituto Federal do Rio Grande do Sul (IFRS), Campus Ibirubá
Ibirubá – RS – Brazil

{cschmitz, galante}@inf.ufrgs.br, {serigne, edimar.manica}@ibiruba.ifrs.edu.br

Abstract. *Ordinances are documents issued by federal institutions that contain, among others, information regarding their staff. These documents are accessible through public repositories that usually do not allow any filter or advanced search on documents' contents. This paper presents ACERPI, an approach which identifies the people mentioned in the ordinances to help the user find the documents of interest. ACERPI combines techniques to discover, obtain, convert and structure documents, extract information, and link employees entities. Experiments were performed on two real datasets and demonstrated a recall of 72.7% for our named entity recognition model trained with only 534 samples and F1 measure of 90% in the efficacy of the entity resolution technique.*

1. Introduction

Brazilian federal institutions publish documents named Ordinances to disseminate changes in their employees' positions, for example, function substitutions, requests for leave, retirement, vacation, among others. Ordinances are official documents issued by organs of the institutions that implement the resolutions contained therein. Today, due to Law no. 12.527 [Brasil 2011], which formalizes the disclosure of information that may be of public interest produced by federal institutions, the publication of the Ordinances by the institutions occurs publicly, allowing anyone to consult them. Access to this information is often given by making the PDF files of the documents available in individual repositories of each institution or even of different campuses within the institutions. With little or no filtering for advanced searches on their content, these document repositories do not allow fast (or even feasible) search for specific employees or types of documents.

Web Scraping techniques have been used to discover and extract files from repositories. This allows fast overcoming of common scraping solutions, such as server-side requests' restrictions. Named Entity Recognition (NER) is applied to identify names in the documents' texts. For this, transfer learning is used to re-train a neural network to the Ordinances domain. Entity Resolution (ER) techniques are then used for matching identified names to real world people, experimenting with different matching criteria. Related approaches include Orion [Manica et al. 2017], which identifies entity-pages for data acquisition; and [Dozier et al. 2010], where Dozier et al. apply different NER techniques, as well as ER in a set of legal documents from the United States of America. ACERPI overcomes the related challenges in a singular approach, combining techniques in a flexible pipeline of data collection, structuring, interpretation and distribution.

This paper proposes an approach, named ACERPI, to discover, obtain, convert and structure files, extract information and solve entities from institutional Ordinances. We collect the documents from their repositories, convert and structure them into a standard XML format, extract relevant information from each of the Ordinances, and resolve the discovered entities. The final result is a document database, flexible and with simplified structures for search of Ordinances and staff members. ACERPI allows the user to search a database using information extracted from the documents, such as the employees mentioned, their identification numbers, the publication date of the Ordinances, and their identification numbers. Experiments with actual data demonstrate the efficacy of the approach, with 98% precision in collection and structuring and an F1 measure of 90% in the efficacy of the entity resolution technique.

The main contribution of this paper is to propose a new approach that combines techniques to discover, obtain, convert and structure documents, extract information, and link employee entities. ACERPI is correctly able to identify people in the ordinances.

This paper is structured as follows. Section 2 reviews related work. Our proposed ACERPI approach is described in Section 3. Section 4 discusses the experimental results while Section 5 concludes the paper.

2. Related Work

Three approaches were deeply related to ACERPI, proposed in this paper. The Orion approach [Manica et al. 2017] aims to discover and extract real entities and attribute values from entity pages. An entity page is a web page that publishes data describing an entity of a given type [Blanco et al. 2008]. Unlike the ACERPI approach, where Natural Language Processing (NLP) is used to identify names in unstructured text, the Orion approach leverages from the structure of the discovered entity pages DOM trees.

In [van Dalen-Oskam et al. 2014], the authors adapted an available NER software to create a Named Entity tagger for Dutch fiction. They also applied Entity Resolution techniques to link the identified Named Entities to Wikipedia entries. They generated a Web Application that provides free-text searching, searching and metadata filtering, and visualization of search results. In ACERPI, queries are available only via database clients, and the creation of a GUI is planned for future work.

Dozier et al. [Dozier et al. 2010] described NER methods using lookup techniques, context rules, and statistical models. They also described techniques employed in resolving entities, such as blocking, features for matching functions, and supervised and semi-supervised learning for the matching function. Furthermore, part of the techniques were used in the extraction and resolution of entities in legal documents from the United States of America, such as jurisprudence cases, depositions, defenses, and other trial documents. The ER technique employed, as opposed to the one used in ACERPI, occurs aiming the association of each entity found to an entry in an authority file. In ACERPI, ER occurs by grouping entities with similar names and contexts.

3. ACERPI

This section describes ACERPI, which is an approach that, by using techniques for file discovery, retrieval, conversion, structuring, information extraction and ER, generates a

database of records and entities which allows searching professional information of public institutions' staff in a categorized, filtered, and clustered manner.

Figure 1 illustrates the data flow from its source to storage and post-processing. ACERPI takes as input a set of documents' repositories. As output, a database is generated with the structured information of the mentioned employees and details of the Ordinances and their metadata. The collection step[1] includes discovering and retrieving the files, converting them to a textual format, and structuring the text documents into XML files, identifying the Ordinances published in the given document. The information extraction step uses Named Entity Recognition [Nadeau and Sekine 2007] and Transfer Learning techniques to identify references to an employee and the related metadata and store them in a standard format. Finally, Entity Resolution techniques [Christophides et al. 2020] are used to relate the identified references to the corresponding real-world personnel and generate the final database. The final database, non-relational and document-oriented (MongoDB), can be used to obtain information about an employee, the ordinances that mention them, and the metadata extracted.

One example is the UFRGS [UFRGS] repository that contains, among other documents, Ordinance 10403 from 11/13/2017, illustrated in Figure 2. This Ordinance indicates a temporary employee replacement and refers to the employees Renata de Matos Galante and Carla Maria dal Sasso Freitas.

3.1. Collection

In this section, we present the strategy to discover, retrieve, convert, and parse the documents into an intermediate, structured format (XML). The initial data, PDF files of the Ordinances, are downloaded from the repositories of the Institutions. The method for **File Discovery and Retrieval** is based on Web Scraping techniques. However, for each type of repository, one or more techniques may be used according to the repository structure and restrictions. Here, we adopted the inference of a navigation pattern [Lage et al. 2004] that, through a regular expression, generalizes the relevant URLs of the repository for automating the retrieval at a later stage.

After discovering and retrieving the PDF files and before starting the extraction step, the **Structuring** sub-step occurs. The structuring goal is to transform the data from its original format (PDF) to an intermediate format with the content of the individual ordinances. Structuring is achieved in two phases: conversion from PDF to text files and interpretation of the content of the files to one or multiple Ordinances.

[1]Developed in partnership with Serigne K. Mbaye and published in his Bachelor Thesis entitled "Developing and Evaluating an Ordinances' Retrieval tool".

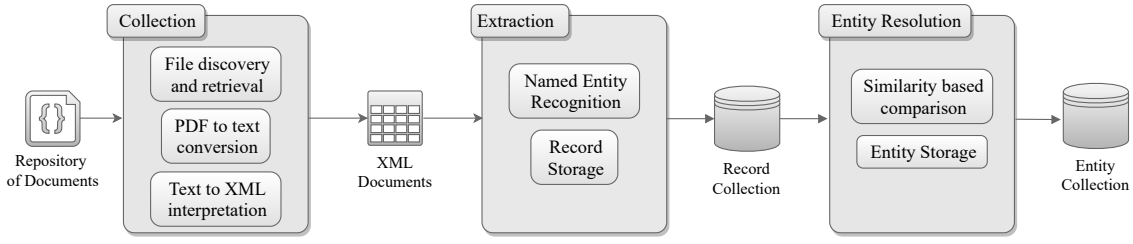
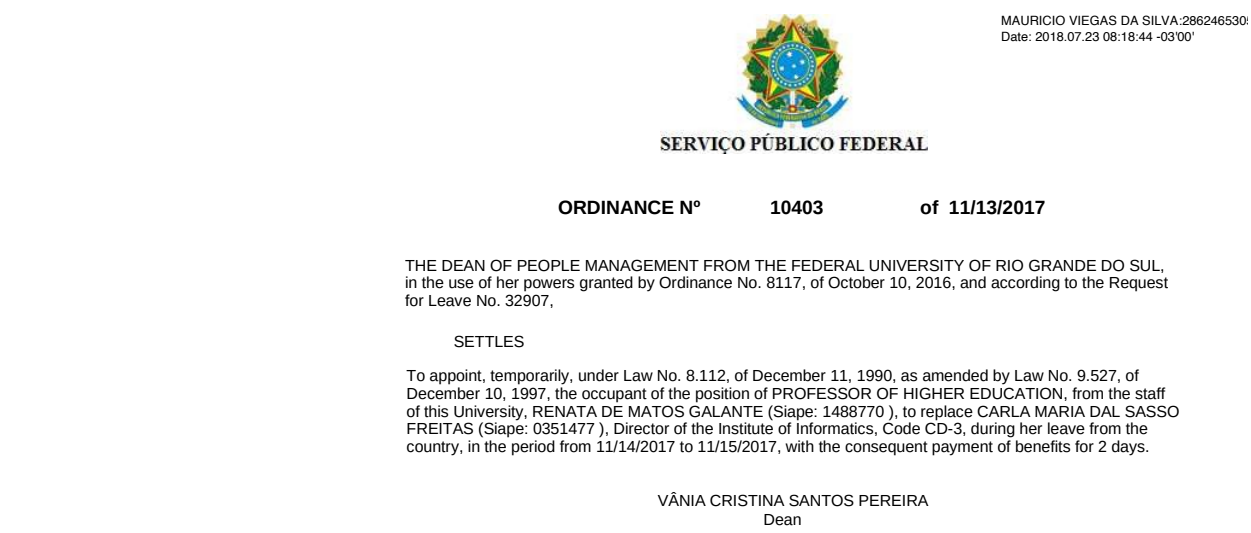


Figure 1. Data flow overview in ACERPI



MAURICIO VIEGAS DA SILVA:286246530
Date: 2018.07.23 08:18:44 -03'00'

SERVIÇO PÚBLICO FEDERAL

ORDINANCE Nº 10403 of 11/13/2017

THE DEAN OF PEOPLE MANAGEMENT FROM THE FEDERAL UNIVERSITY OF RIO GRANDE DO SUL, in the use of her powers granted by Ordinance No. 8117, of October 10, 2016, and according to the Request for Leave No. 32907,

SETTLES

To appoint, temporarily, under Law No. 8.112, of December 11, 1990, as amended by Law No. 9.527, of December 10, 1997, the occupant of the position of PROFESSOR OF HIGHER EDUCATION, from the staff of this University, RENATA DE MATOS GALANTE (Siape: 1488770), to replace CARLA MARIA DAL SASSO FREITAS (Siape: 0351477), Director of the Institute of Informatics, Code CD-3, during her leave from the country, in the period from 11/14/2017 to 11/15/2017, with the consequent payment of benefits for 2 days.

VÂNIA CRISTINA SANTOS PEREIRA
Dean

Figure 2. Ordinance number 10403 of 11/13/2017, issued by the Central Administration of the Federal University of Rio Grande do Sul

First, the conversion of the PDF file to text format is performed. This is achieved by using Apache PDFBox [Foundation], a PDF file manipulation library. Then, in the second phase, the text files' interpretation is performed, extracting the multiple Ordinances that may be contained in each file and their metadata into the intermediate format.

The intermediate format, in XML, has a root element called "document", which has as attributes the unique identifier of the document, the name of the original PDF file, and the file's location in the Institution's repository. A document can have an arbitrary number of children, called "ordinance". The ordinance corresponds to each Ordinance that the original PDF document has, having its number and date as attributes. The pure textual contents referring only and exclusively to the Ordinance identified by the number and date previously extracted are stored within the ordinances elements. Listing 1 shows an example of a structured XML file. The data extraction is performed through regular expressions that capture the character ranges that make up the patterns describing the format of an ordinance, its number, and publication date, respectively.

Listing 1. Intermediate structure parsed from the document displayed in Figure 2

```
<document id="47048" filename="47048.pdf" location="https://wwwl.ufrgs.br/sistemas/
    sde/gerencia-documentos/index.php/publico/ExibirPDF?documento=47048">
    <ordinance nr="10403" date="11/13/2017">
        ORDINANCE No 10403 of 11/13/2017
        THE DEAN OF PEOPLE MANAGEMENT FROM THE FEDERAL UNIVERSITY [...]
        SETTLES
        To appoint, temporarily, [...] the occupant of the position of PROFESSOR OF
            HIGHER EDUCATION, from the staff
        of this University, RENATA DE MATOS GALANTE (Siape: 1488770 ), to replace
            CARLA MARIA DAL SASSO
        FREITAS (Siape: 0351477 ), Director of the Institute of Informatics [...]
        VANIA CRISTINA SANTOS PEREIRA
        Dean
    </ordinance>
</document>
```

3.2. Extraction

This section discusses the techniques used to recognize the names of the employees in the Ordinances and the structure that was chosen to store this relationship.

The first sub-step consists of extracting from the content of the Ordinances the names of the employees mentioned using **Named Entity Recognition**. For this, the natural language processing library Spacy [Explosion.ai a], and a pre-trained model adapted for the Ordinances domain are used. The starting model is `pt_core_news_sm`[2], trained from a news database in Portuguese and convolutional neural networks. With the `pt_core_news_sm` model as a base, data annotation of the Institutions' Ordinances is performed using the Prodigy tool [Explosion.ai b], which provides a Web Application that enables annotation of Named Entities using suggestions provided by the base model. After data annotation, it is necessary to retrain the base model to better interpret the files from the Ordinances domain. This stage is fundamental for improving the recognition of named entities in data with patterns previously unknown by the generic model. This is done again through Prodigy. Given the final model and the ordinance content, the NER output will be the identified Named Entities.

In addition to the name of the employees, context information is extracted from the Ordinances. The SIAPE registration numbers (when present) are extracted via regular expressions. This data corresponds to a unique identification number of the employee, which is used in the ER stage. In Ordinances, these numbers usually accompany the name of the employee, mention the term SIAPE and can be extracted from regular expressions such as `[S|s][I|i][A|a][P|p][E|e][^0-9.]{1,3}([0-9]{6,8})`. If the employee's SIAPE registration number is not precisely identified in the 120 characters following the last character of the employee's name, a list is stored for the employee containing all the registration numbers found in the Ordinance under analysis. This alternative proves useful when names of public servants are mentioned in a list, followed by another list with the respective SIAPE identifiers.

For the PDF file in Figure 2, the output of the extraction stage includes the names of the employees, RENATA DE MATOS GALANTE, CARLA MARIA DAL SASSO FREITAS, and VANIA CRISTINA SANTOS PEREIRA. Their associated SIAPE number, 1488770 for Renata de Matos Galante, 0351477 for Carla Maria Dal Sasso Freitas and both values 1488770 and 0351477 for server Vânia Cristina Santos Pereira. Besides, the association of these data with ordinance 10403 of 13 November 2017 is performed.

Listing 2. Record created for Renata de Matos Galante

```
{
    "id": 131072,
    "name": "RENATA DE MATOS
       GALANTE",
    "siape": ["1488770"],
    "document":
        {"name": "47048"}
}
```

Listing 3. Record created for Carla Maria Dal Sasso Freitas

```
{
    "id": 131073,
    "name": "MARIA DAL SASSO
       FREITAS",
    "siape": ["0351477"],
    "document":
        {"name": "47048"}
}
```

[2]https://spacy.io/models/pt#pt_core_news_sm. Last access in 03/15/2021.

After NER, **Record Storage** occurs. In ACERPI, a main document structure named record was defined, which concentrates the information of a person identified in an ordinance. This database document has, respectively: *(i)* a unique identifier of the record; *(ii)* the name of the individual server identified in the NER step; *(iii)* A list of SIAPEs identified in ordinances related to the employee. This list is populated when it is not possible to identify a specific number for the server, and all the values identified in the document where the respective name was found are inserted into the list; *(iv)* the identifier of the ordinance from which this record was found. Three records are created when analyzing the document in Listing 1. The record with identifier 131072 indicated in Listing 2 of the employee Renata de Matos Galante, the record with identifier 131073 indicated in Listing 3 of the employee Carla Maria Dal Sasso Freitas and a third record in the same basis for the Dean Vânia Cristina Santos Pereira.

3.3. Entity Resolution

Given the records, entity resolution is performed. This step consists of identifying which records refer to the same entities in the real world (i.e., which documents refer to the same staff member, reflecting ordinances in which he/she was directly involved). For example, the records from Listings 2 and 4 both refer to the Professor Renata de Matos Galante. At the end of the ER step, it is expected that the two records are grouped together in the cluster that refers to the real-world entity Renata de Matos Galante, from the Institute of Informatics of the Federal University of Rio Grande do Sul.

ER in the ACERPI approach is performed by grouping the records using **Similarity Based Comparison**. The ER algorithm receives as input the set of records identified from the Ordinances and tries to match each record to any existent group. If no matching occurs, a new group is created containing the record. This loop occurs until all records are grouped and the output is composed of the identified clusters.

The match function is the main part of the algorithm since it defines whether the new record is or is not part of a cluster (i.e., it measures how similar the new record and the records already belonging to a cluster are). It can be simple, as a direct comparison of the named entities of the records, or complex, using methods that compare substrings of the named entities and records' metadata. The ACERPI approach uses a technique that also analyses the context for the resolution of the entities, which in the case of ordinances occurs through the SIAPE identification number. When unique and identical, the SIAPE implies references to the same real-world entity. If the SIAPE numbers are not unique and identical, the comparison of the named entities is performed by cleaning the records' names and comparing them directly. The cleaning procedure consists of characters undercapitalization and trimming. Thus, the records in Listings 2 and 3 would not be grouped because although both have only one SIAPE registration number associated, they differ. On the other hand, the records on listings 2 and 4 would be grouped because they have only one SIAPE registration number each and they are identical.

The clusters resulting from ER proceed to **Entity Storage**. An entity is generated for each cluster, representing a real-world entity. Each entity has a reference to the identifiers of the records that compose it and a set of names and SIAPE registration numbers found in the records to reduce the computational cost of ER. For the records of the Listings 2 and 4, after the step of solving entities, the entity of Listing 5 is generated.

Listing 4. Another record created for the employee Renata de Matos Galante

```
{
    "id": 4630,
    "name": "RENATA DE MATOS
       GALANTE",
    "siape": ["1488770"],
    "document":
        {"name": "50216"}
}
```

Listing 5. Entity generated from entity resolution of the records from Listings 2 and 4

```
{
    "records": [47048, 50216],
    "names": ["RENATA DE MATOS
       GALANTE"],
    "siapes": ["1488770"]
}
```

4. Experimental Evaluation

This section describes two experiments: (1) evaluate data annotation strategies required to correctly extract the named entities from the Ordinances; and (2) experiment and evaluate ER strategies to cluster references to the same staff member.

4.1. Data Sources

Two data sources were used in the experiments:

- **DOCS-UFRGS** Public documents from the Federal University of Rio Grande do Sul. From this source, documents, mostly Ordinances, were extracted from the University's repository. The documents can be accessed through addresses following the navigation pattern [Lage et al. 2004] `https://www1.ufrgs.br/sistemas/sde/gerencia-documentos/index.php/publico/ExibirPDF?documento=[0-9]{1,6}`, where, for this work, the regular expression was replaced by the values 18000 to 105995. For this source, each file contains a maximum of one Ordinance issued by the University. The data was collected until March 3, 2020, with a total storage of 7.99Gb and 44865 PDF files.
- **DOCS-IFRS** Public documents from the Federal Institute of Rio Grande do Sul. For this source, documents were extracted from three repositories, those being: *(i)* Former IFRS - Campus Ibirubá repository [Campus Ibirubá], consisting of files only from Ibirubá Campus from 2013 to 2017(inclusive); *(ii)* Current IFRS - Campus Ibirubá repository [IFRS, Campus Ibirubá], consisting of files only from Ibirubá Campus from 2017 onwards, complementing the previous repository; *(iii)* IFRS general repository [IFRS], consisting documents from all IFRS campi, including ordinances since 2017.

4.2. Experiment 1 - Named Entity Recognition

This experiment evaluates data annotation strategies necessary to correctly extract named entities from Ordinances. In order to do so, training and evaluation of Convolutional Neural Networks (CNN) applied in NER occurs.

4.2.1. Metrics

For this experiment, the metrics used were: *(i)* precision, the percentage of names correctly identified by the model as entity names; *(ii)* recall, the percentage of names that

exist in the entire set that the model was able to identify; and *(iii)* F1-Score. The variables were defined as: *(i)* true positives, the amount of terms identified by the model as an employee name and that actually represented an employee name; *(ii)* false positives, the amount of terms identified by the model as an employee name incorrectly (including names partially found); *(iii)* false negatives, the amount of terms not identified by the model as an employee name but which represented the name of an employee.

4.2.2. Methodology

Annotation sessions and training of models specialized in NER were performed so that they were adapted to understand the structures of the Ordinances. All trainings were based on the pre-existing model `pt_core_news_sm`, trained with news and texts in Portuguese, from which the nuances of the ordinances were transferred through training sessions using annotated data.

The annotation and training process occurs as described in Algorithm 1: data is annotated from annotation suggestions generated by the current model and included in the training set until it is identified that increasing the training set does not bring improvement in the effectiveness of the final model. This is achieved by training the initial model with 25%, 50%, 75% and 100% of the annotated data and calculating the accuracy of the intermediate models[3], indicating (in a simplified way) the improvement of model predictions as the training set increases. When there is no improvement as the training set increases, the model creation process is terminated and the model trained with the data set annotated so far is chosen. From this point on, improvements to the model would require the use of other fine-tuning techniques in models and training, which were not addressed as they do not fit the objective of this work. Both data annotation and training were achieved using Prodigy [Explosion.ai b].

4.2.3. Results

Table 1 shows the metrics extracted during development of the `ufrgs_third` model, used for the DOCS-UFRGS data source. The first row indicates the effectiveness of the `pt_core_news_sm` model when used without training for the recognition of named entities in Ordinances from DOCS-UFRGS, rows 2 and 3 indicate metrics of intermediate

[3]Model evaluation was performed with a set disjoint to the training set.

Algorithm 1: Algorithm used to implement the models.

```
data ← ordinances data set;
annotated_data ← ∅;
initial_model ← pt_core_news_sm;
current_model ← initial_model;
repeat
    annotated_data.add(Annotate(data ∩ annotated_data, current_model));
    current_model ← Train(initial_model, annotated_data);
    Evaluate(current_model);
until accuracy decreases or stabilises with more annotated data;
```

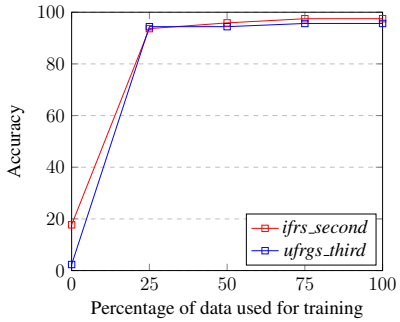


Figure 3. Training curve of the chosen models for each data source

models `ufrgs_first` and `ufrgs_second` and row 4 indicates metrics of the final model used, `ufrgs_third`. In the same way, Table 2 displays the effectiveness of the initial, intermediate and final models trained for the DOCS-IFRS data source. The higher efficacy for the *ifrs* models despite the lower number of examples might indicate overfitting, since the number of employees from IFRS, Campus Ibirubá is small when compared to UFRGS.

The plot in Figure 3 illustrates the stability of accuracy in the training of the definitive models (`ufrgs_third` and `ifrs_second`) with 75% and 100% of the training data, indicating the halt of data annotation. The difference in accuracy between the model `ufrgs_third` without training and trained with only 25% of the annotated data was of 92.07 percentage points. This happens specially since ACERPI leverages transfer learning to achieve decent NER efficacy with a small amount of annotated data.

During results evaluation, some failure cases emerged: *(i)* although the models were trained to identify the complete name of the employees, there were cases where the goal was only partially achieved. As an example, the name of the employee Carla Maria Dal Sasso Freitas was only partially identified as MARIA DAL SASSO FREITAS in some Ordinances; *(ii)* some names were identified as a reference to a single entity, but in fact were referencing many. This may happen due to the structuring of the Ordinances in previous steps, which may end up removing line breaks sometimes used as delimiters of the end of a name in a list. As an example, one term identified as a reference to a single entity "NAIRA MARIA BALZARETTI RENATA JENISCH BARBOSA TANIRA" actually refers to three entities: Naira Maria Balzaretti, Renata Jenisch Barbosa and Tanira Rodrigues Soares; *(iii)* identification of service provider company names. This may happen due to the model learning to identify the employees as objects in the sentences, which ends up misleading it to induce that the object of a sentence of an Ordinance is always a person name, when in fact it is not. As an example, the term "MEGATRON ENGENHARIA LTDA" was identified as a named entity in several documents.

4.3. Experiment 2 - Entity Resolution

This experiment evaluates resolution strategies from named entities to real-world entities. Thus, the focus of the experiment is to evaluate the effectiveness of the proposed approach to group all mentions of an employee identified in the most diverse retrieved documents.

Table 1. Effectiveness of the NER models for DOCS-UFRGS

Model	Precision	Recall	F-1	Examples
pt_core_news_sm	1,0%	10,4%	1,9%	0
ufrgs_first	64,7%	71,4%	67,9%	100
ufrgs_second	65,1%	72,7%	68,7%	204
ufrgs_third	70,0%	72,7%	71,3%	534

Table 2. Effectiveness of the NER models for DOCS-IFRS

Model	Precision	Recall	F1	# Examples
pt_core_news_sm	11,1%	41,1%	17,5%	0
ifrs_first	85,2%	87,8%	86,5%	204
ifrs_second	87,9%	92,2%	90,0%	418

4.3.1. Metrics

The metrics chosen to evaluate the ER are based on pairwise comparison, where each pair indicates a relationship between entities (in order to refer to the same entity in the real world). For a grouping of entities $A = \{entityX, entityY, entityZ\}$ the pairs $A_p = \{\{entityX, entityY\}, \{entityX, entityZ\}, \{entityY, entityZ\}\}$ are generated and, from those pairs, the evaluation metrics are calculated. The metrics used for this experiment were: *(i)* precision, the percentage of identified pairs that actually refer to the same entity in the real world; *(ii)* recall, the percentage of all pairs referring to the same entity that were correctly identified; and *(iii)* F1-Score. The variables were defined as: *(i)* true positives, the pairs that actually refer to the same entity in the real world; *(ii)* false positives, the pairs that were identified as references to the same entity in the real world, but are not; and *(iii)* false negatives, the pairs that were not identified as references to the same entity in the real world, but are.

4.3.2. Methodology

For this experiment, were solved 194 thousand records found in the Ordinances from the DOCS-UFRGS data source using different match functions in order to identify the impact of the function variations in the effectiveness of the ER process. The chosen match functions were: (1) string comparison of the names identified in the given records; (2) string comparison of the names present in the records with cleansing for removal of blank sequences and name uncapitalisation; and (3) the presence, in both registers, of only one SIAPE enrollee and their identity. In case there is only one SIAPE value in each register but they are different, the match function returns false. If there are multiple SIAPE numbers in at least one record, the SIAPE plate is ignored and the matching occurs according to the previous item. For evaluation, all references to the 24 titular professors of the Institute of Informatics of the Federal University of Rio Grande do Sul were manually discovered and clustered.

Table 3. Entity resolution matching functions' efficacy evaluation.

Matching Function	Precision	Recall	F1
1 - Original names	100,0%	59,2%	74,4%
2 - Pre-processed names	100,0%	75,5%	86,0%
3 - SIAPE values + pre-processed names	99,5%	82,4%	90,1%

4.3.3. Results

Table 3 indicates the results of the experiments for the different matching functions. It can be observed that the precision was kept close to 100% and decreased as the match function complexity increased. This happens because the used functions were conservative, so that they only grouped records with the same name or equal names but with different formatting (amount of spaces and/or capitalization). As the match function becomes more complex, the precision tends to decrease in detriment of the revocation, which increases due to the flexibility of the clustering conditions.

The selected approach was number 3, with an F1 measure of 90.1%, it also demonstrated a 6.9% improvement in recall, without drastically affecting precision, which was reduced by only 0.5%. This reduction could be a side-effect of trusting that when both records contain one SIAPE value each and they match, the records refer to the same real world entity, which might not be the case since we also attach to records all the SIAPE numbers found in the document when none is found in the following 120 characters.

During results evaluation, some failure cases emerged: *(i)* records were identified containing partial names or prefixes and suffixes which were not grouped together with the ones with the correct and complete names. As an example, the records with name "LUCIANE MACHADO CAETANO MOSSMANN-", were not grouped with those with name "LUCIANE MACHADO CAETANO MOSSMANN"; *(ii)* names spelled differently but referred to the same entity in the real world. Since the documents under analysis are written by different people and are susceptible to writing errors, it may happen that the same name is written in different ways, as is the case of the staff member Sérgio Bampi, who has references in documents both as "SÉRGIO BAMPI" and as "SERGIO BAMPI". These records were not grouped together.

5. Conclusion

In this paper, an approach for processing Ordinances from federal institutions was presented, which, in the end, generates a non-relational database for querying staff and Ordinances. The effectiveness of our approach was proven with our experiments on two real data sources, one with over 40 thousand files and thousands of employees mentioned. Results demonstrated the efficacy of the approach, with an F1 measure of 90% in the efficacy of the entity resolution technique. The main contribution of this paper is the creation of a flexible Ordinance data processing pipeline, with independent intermediate steps, with the definition of extensible intermediate structures, explaining the choices and tools used, and evaluating the implementation on a data source. Main limitations include institution-specific NER tagger, relatively simple similarity functions in ER and lack of a GUI for user search and browsing. For future work, we intend: *(i)* the use of a blocking technique to reduce the number of comparisons needed during ER and allow more complex match-

ing functions; *(ii)* a graphic interface for advanced search regarding staff and Ordinances; *(iii)* The classification of the content of the Ordinances into known categories.

References

[Blanco et al. 2008] Blanco, L., Crescenzi, V., Merialdo, P., and Papotti, P. (2008). Supporting the automatic construction of entity aware search engines. In *Proc. of the 10th ACM Workshop on WIDM*, page 149–156, NY, USA.

[Brasil 2011] Brasil (2011). Lei nº 12.527/2011. *Diário Oficial da República*.

[Campus Ibirubá] Campus Ibirubá. Boletins de Serviço. `https://ibiruba.ifrs.edu.br/site/conteudo.php?cat=50`. Accessed: 2021-09-05.

[Christophides et al. 2020] Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., and Stefanidis, K. (2020). An overview of end-to-end entity resolution for big data. *ACM Comput. Surv.*, 53(6).

[Dozier et al. 2010] Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., and Wudali, R. (2010). *Named Entity Recognition and Resolution in Legal Text*. Springer-Verlag.

[Explosion.ai a] Explosion.ai. Industrial-strength natural language processing. `https://spacy.io/`. Accessed: 2021-09-05.

[Explosion.ai b] Explosion.ai. Prodigy · radically efficient machine teaching. an annotation tool powered by active learning. `https://prodi.gy/`. Accessed: 2021-09-05.

[Foundation] Foundation, T. A. S. Apache pdfbox® - a java pdf library. `https://pdfbox.apache.org/`. Accessed: 2021-09-05.

[IFRS] IFRS. Documentos. `https://ibiruba.ifrs.edu.br/site/conteudo.php?cat=50`. Accessed: 2021-09-05.

[IFRS, Campus Ibirubá] IFRS, Campus Ibirubá. Boletim de Serviço. `https://ifrs.edu.br/ibiruba/documentosoficiais/boletim-de-servico/`. Accessed: 2021-09-05.

[Lage et al. 2004] Lage, J. P., Silva, A. S., Golgher, P. B., and Laender, A. H. F. (2004). Automatic generation of agents for collecting hidden web pages for data extraction. *DKE*, 49:177–196.

[Manica et al. 2017] Manica, E., Dorneles, C. F., and Galante, R. (2017). Orion: A cypher-based web data extractor. In *DEXA*, pages 275–289, Cham. Springer.

[Nadeau and Sekine 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26.

[UFRGS] UFRGS. Consulta a portarias geradas pela reitoria da ufrgs. `https://www1.ufrgs.br/sistemas/sde/gerencia-documentos/index.php/publico/consultar/`. Accessed: 2021-09-05.

[van Dalen-Oskam et al. 2014] van Dalen-Oskam, K., de Does, J., Marx, M., Sijaranamual, I., Depuydt, K., Verheij, B., and Geirnaert, V. (2014). Named entity recognition and resolution for literary studies. *Computational Linguistics in the Netherlands Journal*, 4:121–136.