Thomas Zaragoza, Yann Nicolas and Aline Le Provost

17 From Text to Data Inside Bibliographic Records: Entity Recognition and Linking

Abstract: The *Système Universitario de Documentation* (Sudoc)/University Documentation System catalogue is the French higher education union catalogue. It is run by *Agence bibliographique de l'enseignement supérieur*/Bibliographic Agency for Higher Educations (Abes). Like any large database, Sudoc with 15 million records has quality issues that can negatively impact the user experience or database maintenance efforts, for example the process towards an IFLA Library Reference Model (LRM) compliant catalogue. Quality issues are diverse: data might be inaccurate, ambiguous, miscategorised, redundant, inconsistent or missing. Details might not really be missing but hidden or lost in text inside the bibliographic record itself. This chapter describes efforts to extract structured information about contributors and their roles from statements of responsibility to generate automatically data in access points: last name, first name, relator code and optionally identifier, to link to the French higher education authority files. The first step in the process was a named entity recognition task implemented through a machine learning (ML) approach. The second step was an entity linking task. The pipeline is for Abes a first experience in adopting machine learning and building a generic approach to AI uses in cataloguing.

Keywords: Cataloguing; Entity-relationship modelling; Machine learning; Text analysis (Data mining)

Introduction

The *Système Universitaire de Documentation* (Sudoc) catalogue is the union catalogue of French academic and research libraries and it is managed and maintained by *Agence bibliographique de l'enseignement supérieur* (Abes)/Bibliographic Agency for Higher Education. Library catalogues are old, large and highly structured databases, created and maintained by professionals with a strong quality ethos. The development of new information technology paradigms has stressed the importance of data quality. On a web of linked data, the pollution of good data by less good data is a permanent risk and predictive models and decision making algorithms require the best possible training data to optimise results and minimise the generation of additional erroneous data.

https://doi.org/10.1515/9783111336435-018

As in any large database, quality issues in Sudoc's 15 million bibliographic records are diverse: data may be inaccurate, ambiguous, miscategorised, redundant, inconsistent or missing. Data are not necessarily missing; they might be implicit, hidden or lost in text inside the bibliographic record itself. Inaccurate or incomplete data impinge on the usefulness of the catalogue to its users and also inhibit future maintenance and moves to new approaches and conceptual models like the IFLA Library Reference Model (LRM) (Riva, Le Boeuf, and Žumer 2018).

This chapter focuses on one important and interesting aspect of so-called missing data. Contributor names and roles are transcribed from a document to the appropriate MARC descriptive field statement of responsibility, referred to throughout this chapter as the SoR. Most names and roles have corresponding access points that contain the normalised name and a function relator code to express the role, optionally the identifier of an authority record. But in Sudoc, many records have contributor mentions in descriptive fields that are not identified in access points. There are 300,000 existing person access points which lack a relator code.

This chapter describes the current effort being undertaken to extract structured information or data about contributors and their roles from text in the SoRs. The objective is to generate automatically, or correct, access points containing the last name, first name, function relator code, and optionally identifier, to link to the authority files maintained by Abes and used by the French higher education catalogue Sudoc. The authority files known as *Identifiants et Référentiels pour l'Enseignement supérieur et la Recherche* (IdRef)/Identifiers and Repositories for Higher Education and Research (Abes n.d. a) provide a variety of records within its database:

> An open and reusable database, IdRef provides more than 6 million authority records of different types:
> - Nearly 4 million individuals
> - More than 400,000 local authorities (legal entities) and congresses
> - More than 120,000 geographic locations
> - ...
>
> In the documentation ecosystem, authority records are used to control certain information that is common to several bibliographic records. They are used to identify, describe, aggregate, bounce back and disambiguate the resource being described.
>
> The richness of authority records is also expressed in the quality and completeness of the links that unite them to bibliographic records, which makes it easy to establish an author's bibliography (Abes n.d.b).

The work being undertaken has two parts. The first is the creation and evaluation of a named entity recognition (NER) model to extract person and role entities from the SoRs. The second is linking the extracted role to the UNIMARC controlled

vocabulary of roles via text classification. A pre-existing generic model has been employed for the recognition of names and retrained with ad hoc data marked up by librarians through a dedicated annotation tool. For the extraction of roles, a model was generated from scratch. The linking of contributor names has been achieved using a logical rule based artificial intelligence (AI) framework which is currently still being debated with a preference for either an entity linking model or a classification model over a rule-based approach. This project is a work in progress and the last section of the chapter provides an overview of further work to be done.

Incomplete Bibliographic Records

For various reasons, many bibliographic records contain incorrect and/or incomplete data. Some examples are provided of such records. The following UNIMARC bibliographic record has as its statement of responsibility (SoR):

200 1 **$a**Hawking**$f**Stephen Finnigan, réalisation**$g**Stephen Hawking, Stephen Finnigan, Ben Bowie, scénario**$g**Joe Lovell ; Tina Lovell ; Arthur Pelling [et al.] acteurs

B200$f and B200$g UNIMARC fields encode the SoR as found on the document with the title frame as a source for a motion picture or the title page for a book. The SoR transcribes and records the original text found on the document, with minimal structuring and a separate SoR per role or function. A unique SoR can mention more than one person, for example:

$gStephen Hawking, Stephen Finnigan, Ben Bowie, scénario

The following UNIMARC fields deal with access, not description. This highly structured data comes from the intellectual analysis of the document by the cataloguer, not the transcription of the title page:

700 1 **$3**241177782 **$a**Finnigan **$b**Stephen **$4**300
701 1 **$3**028590295 **$a**Hawking **$b**Stephen **$4**690
701 1 **$3**241177286 **$a**Lovell **$b**Joe **$4**005
701 1 **$3**241177421 **$a**Lovell **$b**Tina **$4**005
701 1 **$3**241177588 **$a**Pelling **$b**Arthur **$4**005

Each line is called an access point and refers to a unique person. In 701 1 $3241177421 $aLovell $bTina $4005, the subfields respectively refer to the linked authority record (Abes n.d.b) the last name, the first name and the UNIMARC role

code (IFLA 2021) of the person related to this document. It is easy to peruse the previous SoRs and count the number of different persons mentioned, that is six, and then compare this number with the number of access points, that is five and come to the conclusion that the record lacks access points. But this conclusion is not trivial to reach automatically. In the process of encoding the five access points, the cataloguer extracted person names from the three SoR fields, but seems to have dismissed or forgotten one person, Ben Bowie. The project intention is to rely on machines to extract the named person missing, and to predict the precise role or function Ben Bowie plays in the work.

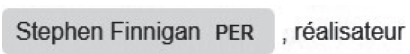
There are 300,000 author access points which lack a relator code and constitute lacunae that the ML project will try to fill. Missing access points in the Sudoc database are more difficult to find and require knowing how many different names are mentioned in the SoR. That is precisely one of the objectives of the project and can be accomplished using a named entity recognition model.

Named Entity Recognition and Extraction

Named Entity Recognition (NER) is an application of Natural Language Processing (NLP) on large amounts of unstructured text to extract entities. The application chosen for the purpose was an open-source model offered by spaCy. The model used was fr_core_news_lg, that had been trained on 7200 high quality, hand annotated French articles from Wikipedia, WikiNER (Nothman et al. 2013), and over 3000 French sentences from UD French Sequoia v2.8 (Candito et al. 2014). It can be used to extract persons, organisations and locations but not roles. For the latter, a new model was created and trained from scratch.

When applying the initial model on the statement of responsibility:

Stephen Finnigan, réalisateur, the result is:



And whilst offering promising outcomes in a broad context, the result does not translate well into a bibliographic context. Statements of responsibility are not written as naturally as the articles and sentences on which the model was trained. The model also ignores the role entity. Retraining of the model in a bibliographical context would be required using a large quantity of annotated bibliographic records. To accomplish this task, the annotation tool Prodigy was used which facilitated speedy and easy annotation of a large amount of data with spaCy's models built-in for training and retraining. SpaCy's capacity for recognising people can be

retained, but it has no capacity to recognise roles. After annotating 10,000 bibliographical records using Prodigy, spaCy's model was retrained to extract person entities and to begin extracting role entities.

Evaluation

A new model can be evaluated using three metrics: precision, recall and accuracy. The metrics can be visualised using a confusion matrix (Figure 17.1).

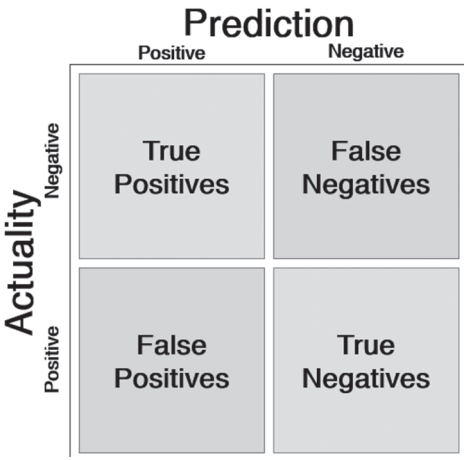


Figure 17.1: Confusion matrix example

Precision is the chance for a prediction to be true and is calculated by dividing the number of true positives by the sum of true positives and false positives. Recall is the chance for an entity of a certain class to be predicted as such. It is calculated by dividing the number of true positives by the sum of true positives and false negatives. Finally, the overall accuracy of a model is the sum of true positives and true negatives divided by the sum of all four.

To return to the previous problem, after retraining the model, the results shown in Figure 17.2 were obtained:

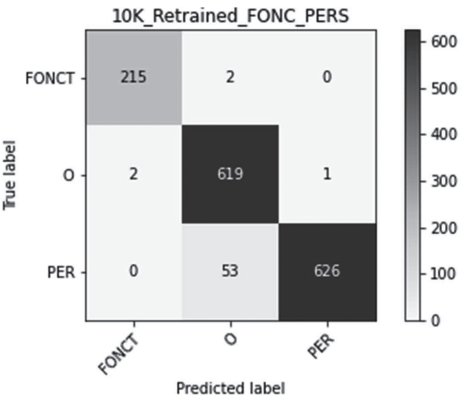


Figure 17.2. Confusion matrix of retrained spaCy model

Table 17.1: Results with retrained spaCy model

	Precision	Recall	F1-Score	Support
Fonct (Role)	0,99	0,99	0,99	217
O (Outside)	0,92	0,99	0,96	622
Per (Person)	0,99	0,92	0,96	679

The retrained model provides good performance (Table 17.1) and was able to detect 99% of person and role or fonct entities of which 99% of its predictions were correct. In the table, O indicates outside of an entity, and it is a token or a word that is neither role nor person. The model proved sufficiently efficient for tackling the next objective while still being able to further retrain the model, given additional annotated data. To reiterate, a suitable NER model was available to analyse statements of responsibility to detect the people mentioned and the keywords representing their roles in the creation of the catalogued document.

Linking Role Keywords

The next step in the process was to link role keywords to their controlled role codes. The example previously used illustrates the issue:

> ***Stephen Finnigan, réalisateur***

There are approximately 155 different relator codes. For a human, it is quite easy to link *réalisateur*/director in the context of a movie to the function relator code

300 - Movie director. But there can be difficulties clearing ambiguity without sufficient context. A director could also be 632 - Artistic director or 727 - Thesis Director. To automatically remove the ambiguity, the first approach was an entity-linking model, which links entities of a text to a controlled ontology. The problem encountered with this approach was that it relied on a rules-based system, which could produce excellent results, but also required a larger quantity of effort which lacked viability within the project context. Instead, a text classification approach was chosen.

Text classification is a machine learning technique that assigns a set of predefined categories to open-ended text through variables like word count frequencies and other predefined features and requires no set of rules. The human work is limited to choosing features, extracting data, creating training sets, and choosing an algorithm that provides the best result for our problem. It was decided to provide the following features initially: role keywords with the entities extracted by the NER model including directed, written or illustrated; the document content type from a list including text, still image, animated image, and video; the position of the mention of responsibility relative to other SoRs in the same record; and the document type in relation to its being a thesis or academic paper, or not.

Creation of the Training Dataset

To obtain the features outlined and create a training set, the entries for the UNIMARC fields 200$ (SoR), 70X$ (access points), 181$c (content type), 608$3 (ID of the authority record of the form/genre of the document), 105$a (textual resources types) and 503$a (form title) were extracted from the Sudoc database. With the NER model, the role keywords and the person names were extracted from the SoR and subsequently paired with the persons in the access points to provide a relator code answer for the training of the model. Following extraction of the raw data, the position of each SoR was noted, including whether the document was a thesis or not. Table 17.2 outlines the entries obtained. The content type, tdi, refers to biodimensional animated image.

Table 17.2: Extraction of data

Bibliographic Record ID	SoR	Position	Access Points	Content type	Thesis
236018256	Stephen Finnigan, director	0	Finnigan Stephen 300 Hawking Stephen 690 Lovell Tina 005 Lovell Joe 005 Pelling Arthur 005	tdi	False
236018256	Stephen Hawking, Stephen Finnigan, Ben Bowie, scénario;	1	Finnigan Stephen 300 Hawking Stephen 690 Lovell Tina 005 Lovell Joe 005 Pelling Arthur 005	tdi	False
236018256	Joe Lovell, Tina Lovell, Arthur Pelling [and al.] auteurs	2	Finnigan Stephen 300 Hawking Stephen 690 Lovell Tina 005 Lovell Joe 005 Pelling Arthur 005	tdi	False

The most complex step is executing the NER model on the SoRs to extract persons and roles keywords. The keywords are linked to the appropriate persons using a simple Person* + Role* / Role* + Person* pattern (Figure 17.3):

Figure 17.3: Statement of responsibility example pattern link

A string comparison of the person entity was made with the persons in the access points taking in consideration abbreviations, composite names and typos using the Levenstein distances between the two names. Given that the NER model has correctly extracted the entities, and that the extracted name and access point name have been paired, a training set can be created (Table 17.3).

Table 17.3: Training set example

Keywords	Position	Content type	Thesis	Label
réalisateur	0	tdi	False	300
scénario	1	tdi	False	690
acteurs	2	tdi	False	5

The approach has its limits. Since it is vital that the training set is as accurate as possible any ambiguity encountered that could not be removed was ignored. For example, the correctly completed bibliographic record, Stephen Finnigan, would have the following roles: 300 and 690. And in such case, it would be impossible to distinguish between the associated keywords and function relator codes in the entry (Table 17.4).

Table 17.4: Ambiguous bibliographic record example

Bibliographic Record ID	Keywords	Position	Content type	Thesis	Label
236018256	Directeur, scénario	0,1	tdi	False	300,69

Directeur could be associated with the role code 300 or 690, and it requires a choice between the two which is the objective of the model. It poses a conundrum as the model's predictions cannot be used to create its own training set. At least not yet.

The inability to differentiate between multiple role codes is limiting and potentially discriminatory in the training of the model as some roles have a greater chance of appearing in conjunction with others than alone and tend to have fewer training entries. Nonetheless a training set was constructed with 1000 entries per function for thirteen of the most common functions: 005 – Actor , 065 – Auctioneer, 070 – Author, 100 – Original Author, 230 – Composer, 300 – Movie director, 340 – Scientific publisher, 365 – Expert, 440 – Illustrator, 651 – Publication director, 727 – Thesis director, and 730 – Translator. A limit of 1000 entries was set. Whilst there are hundreds of thousands of annotated training entries of certain classes like the role code 070 – Author, there are only thousands of entries of minority classes like the role code 300 – Composer. A balanced training set must be created to avoid bias in a predictive model. The easiest and safest method to do so was to downscale, which entailed creating a training set with an equal number of entries per role code. The number chosen must be the highest possible while remaining lower than

the number of training entries from the lowest count class. Once the training set was created, it was then used to train a k-nearest neighbours algorithm model.

K-Nearest Neighbours Algorithm (KNN)

"If it looks like a duck, if it sounds like a duck, then it probably is a duck."

The k-nearest neighbours algorithm (KNN) was implemented initially to evaluate the feasibility of the approach before branching out and experimenting with various different algorithms to choose the most appropriate, because the KNN algorithm is simple to comprehend and constitutes a simple introduction to the world of machine learning and decision-making algorithms.

The entries in the training set are encoded into numerical values that can be computed by the model. Each training entry constitutes a point in a universe. For example, if a KNN model is trained with k=3 and two classes, author in red and illustrator in blue, any attempt to predict the classification of a new point, in this case grey, will look for the three nearest neighbours and predict the majority, in this case, the role of illustrator in blue (Figure 17.4).

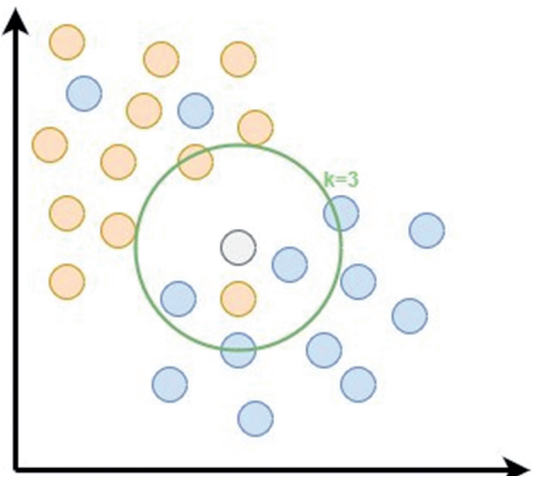


Figure 17.4: Prediction by KNN model with k=3

Evaluation

After training the model, the results can be visualised in Table 17.5, and in the confusion matrix (Figure 17. 5).

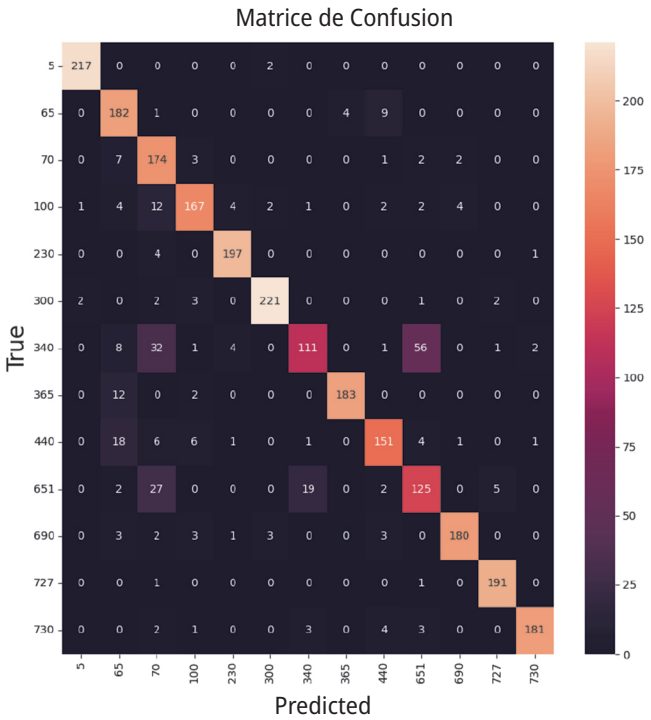


Figure 17.5: Confusion matrix for KNN model, K=6 , 13k training

Table 17.5: Evaluation table

Label	Precision	Recall	F1-Score	Support
005	0,99	0,97	0,98	224
065	0,93	0,78	0,85	211
070	0,44	0,91	0,6	196
100	0,96	0,81	0,88	217
230	0,99	0,98	0,98	179
300	0,92	0,98	0,95	191
340	0,95	0,61	0,74	201
365	0,97	0,9	0,93	181
440	0,96	0,78	0,86	197
651	0,81	0,6	0,69	58

Label	Precision	Recall	F1-Score	Support
690	0,94	0,96	0,95	206
727	0,95	0,99	0,97	207
730	0,99	0,89	0,94	205

The greatest discrepancy is in the 070 - Author column of predictions. This class has a poor precision measurement of 0.444 which is due to the nature of the class. Being the most common and principal role in the Sudoc, often in first position. and not introduced by a role keyword, the conclusion is implicitly reached that the said person is the author. The lack of role keywords makes the author class a collective bin for other entries that lack role keywords due to either a mistake by the NER model or initial cataloguing, or differences in the average entry of its class. In practice since the class 070 – Author is a majority role (Figure 17.6), the fact that it is the de facto prediction for sparse entries is the least damaging to overall efficiency. The satisfactory recall and precision scores for other classes lend assurance to accurate prediction for more specific and least common classes. Overall, the model appeared to be satisfactory, particularly given the limited amount of data used to train and test it.

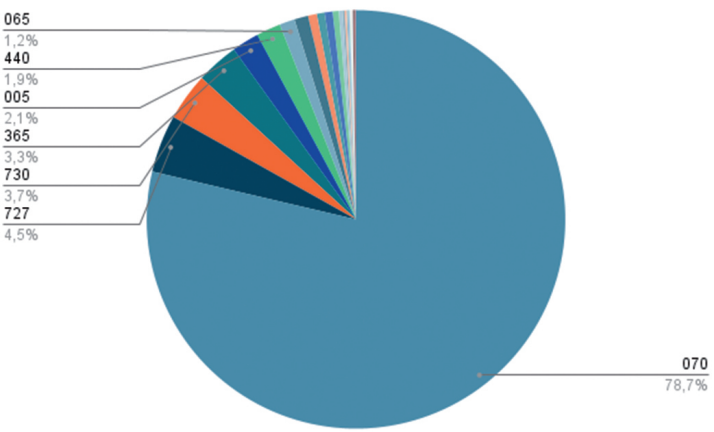


Figure 17.6: Pie chart of the distribution of function relator codes

As previously mentioned, there are over 100 roles that the model must differentiate which is considerably more than the thirteen roles in the training set. The current approach may well be flawed. Another model was trained, this time differentiating between thirty-five classes, and different results emerged (Table 17.6).

Table 17.6: Evaluation table with 35 classes

Label	Precision	Recall	F1-Score	Support
003	0.67	0.01	0,03	114
005	0.9	0.99	0,94	192
010	0.84	0.69	0,76	146
040	0.15	0.24	0,18	169
065	0,81	0,78	0,79	185
070	0.37	0,9	0,53	194
080	0,84	0,81	0,83	199
100	0,74	0,74	0,74	196
180	0,96	0,93	0,95	222
205	0.5	0.78	0,61	218
212	0.82	0.8	0,81	175
220	0.76	0.59	0,66	136
230	0.91	0.94	0,93	203
273	0.47	0.44	0,45	213
300	0.92	0.95	0,94	198
340	0.67	0.35	0,46	204
350	0.62	0.86	0,72	196
365	0.96	0.88	0,92	199
410	0.93	0.87	0,9	190
440	0.83	0.62	0,71	213
460	0.76	0.64	0,69	181
470	0.71	0.37	0,49	147
550	0.92	0.88	0,9	132
555	0.38	0.08	0,14	173
595	0.53	0.25	0,34	201
600	0.97	0.72	0,83	199
651	0.61	0.47	0,53	185
673	0.21	0.86	0,34	217
690	0.97	0.86	0,91	222
710	0.88	0.59	0,7	192
727	0.85	0.08	0,16	198
730	0.99	0.88	0,93	210
956	0.88	0.51	0,65	192
958	0.82	0.48	0,61	205

The cells with a score lower than 0.7 have been highlighted and it can be inferred that there are many more collective bins, due in part to the similarities between roles that did not exist in the previous model. Figure 17.7 charts and simplifies the mispredictions.

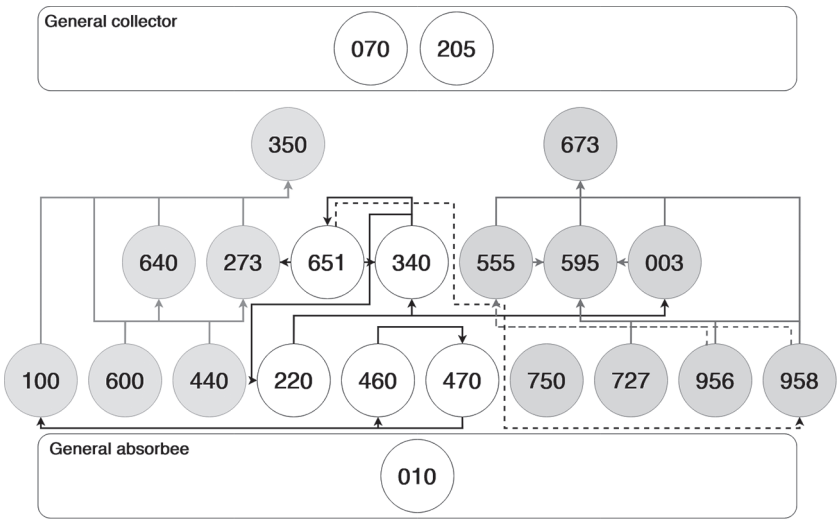


Figure 17.7: Prediction graph of function relator codes

Two specialised collective bins for classes 350 – Engraver, and 673 – Director of the research team, emerged representing respectively the artistic and academic classes. The existence of the collective bins indicates insufficient features to distinguish satisfactorily between similar roles. Further diagnostic work is required to determine if and how the results can be improved.

Further Work

The next stage of the work is to optimise and improve on the results achieved and work accomplished. There are over one hundred function relator codes. The text classification model implemented was able to predict only thirty-three function relator codes. And the results obtained justified an eventual implementation of the models on a grand scale for only thirteen function relator codes among the thirty-three function relator codes. To improve results, it is planned to vastly increase the amount of data allocated to the text classification model's training. To overcome the current constraint of keeping the classes balanced, the number of bibliographical records extracted and used to automatically create training entries will be significantly increased and/or records will be manually extracted including the lacking function relator codes to create specific training entries.

It is also planned to increase the number of features on which a model may rely when making its decisions to enhance the quality of the prediction and to differen-

tiate more accurately between similar function relator codes. The technical team is working closely with a team of cataloguing experts to analyse the false predictions and explore potential new features. Another possible solution is the creation of different models for different needs. Multiple specialised models targeting specific categories of documents could be created and trained rather than using a large all-encompassing model. Class constraints in creating balanced training sets would be reduced and differentiated models would be able to discriminate between similar role codes through focusing on finer details and differences. Another possibility is the exploration and implementation of different classification algorithms apart from KNN to find one better suited to the requirements.

The implementation of an additional decision-making algorithm above the current prediction model would increase the accuracy. The aim would be to analyse the score given by the initial prediction model and act upon it only at a certain threshold. To present this idea in simpler terms, the prediction would be taken into consideration only if it were sure enough and ignored if deemed unsure. Precision would be enhanced at the cost of recall, which would be a sacrifice worth making given the nature of the project. There would also be the option of accepting multiple predictions instead of one. The hierarchy between roles should also be taken into consideration. For example, 350-engraver is a more specific type of author 070. Finally, the generation of missing access points must be implemented. Differentiation between first and last names to link the person names to their idRef identifier could be undertaken using the logical rules-based system tool Qualinka (Le Provost 2020). If the linking is not possible, first and last names will still need to be differentiated when creating UNIMARC zones and subzones for access points.

Conclusion

The objective of the project was to teach models to analyse the text of statement of responsibility to automatically generate missing access points for contributors or add the function relator code when missing in an existing access point. It first required the extraction of two types of entities: persons and roles. The standard Spacy model used for the task gave excellent results with a precision > 0.99 and recall > 0.92 after being re-trained on 10,000 manually annotated records. The annotation task, critical but often time-consuming, was efficiently achieved with the help of a dedicated annotation web tool. Librarians can use AI to produce more reliable training data which will feedback into further improvement of the AI.

When the persons and roles are extracted and paired, the next arduous task is to find the right function relator code among UNIMARC relators. Our first results

with KNN as classification algorithm will be completed and hopefully enhanced in various ways: more data, more features, more specialised models, more algorithms; and more librarians to analyse the predictions and work with the data scientist.

This project is still a work in progress. If the end result is satisfactory, it could be used in production to create new access points for bibliographic records or check if the existing ones are correct. But no matter what the final conclusion of the project might be, it will have taught Abes a great deal both as a bibliographic agency and as a data steward of massive quantities of data. Abes has a duty to understand and adopt machine learning approaches as quickly as possible to fulfil its traditional missions. Abes with the help of efficient tools to annotate and prepare the data, and with librarians in the loop (Lieber, Van Camp, and Lowagie 2022), can achieve real progress both in terms of data quality and human resource development, two main issues for libraries.

Acknowledgments

The authors are grateful to Pascal Poncelet, University Montpellier, Abes colleagues who helped us to extract and analyse data, and Stephen Finnigan.

References

Agence bibliographique de l'enseignement supérieur (Abes). n.d.a. "IdRef: Identifiants et Référentiels pour l'Enseignement supérieur et la Recherche/Identifiers and Repositories for Higher Education and Research." https://www.idref.fr/autorites.jsp.

Agence bibliographique de l'enseignement supérieur (Abes). n.d.b, "Données d'autorité et référentiels." https://abes.fr/reseaux-idref-orcid/le-reseau/.

Candito, Marie, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah and Éric de la Clergerie. 2014. "Deep Syntax Annotation of the Sequoia French Treebank." In Proceedings of Language Resources and Evaluation Conference 2014, Reykjavík, Iceland. https://inria.hal.science/file/index/docid/971574/filename/deep_sequoia.final_with_keywords.pdf. Updated dataset used posted 2023 https://github.com/UniversalDependencies/UD_French-Sequoia?tab=readme-ov-file

International Federation of Library Associations and Institutions (IFLA). 2021. "Appendix B: Relator Codes." In *UNIMARC Bibliographic Format Manual (online ed., 1.0, 2021).* https://www.ifla.org/wp-content/uploads/2019/05/assets/uca/unimarc_updates/BIBLIOGRAPHIC/u_b_appb_update2020_online_final.pdf.

Le Provost, Aline, and Yann Nicolas. 2020. "IdRef, Paprika and Qualinka**.** A Toolbox for Authority Data Quality and Interoperability**"** *ABI Technik* 40, no. 2 : 158–168. https://doi.org/10.1515/

abitech-2020-2006. Available at https://abes.fr/en/publications/articles-et-contributions/idref-paprika-and-qualinka/.

Lieber, Sven, Ann Van Camp, and Hannes Lowagie. 2022. "A LITL More Quality: Improving Library Catalogs with a Librarian-in-the-loop Linked Data Workflow." Semantic Web in Libraries (SWIB) Conference, November 28, 2022. Presentation by KBR, Royal Library of Belgium. *YouTube.* Video. 27.31. https://www.youtube.com/watch?v=r29W73vle2I

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy and James R. Curran. 2013. "Learning Multilingual Named Entity Recognition from Wikipedia." *Artificial Intelligence* 194: 151-175. https://doi.org/10.1016/j.artint.2012.03.006. Dataset used posted October 3, 2017 at https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500.

Riva, Pat, Patrick Le Boeuf, and Majal Žumer. 2018. "IFLA Library Reference Model: A Conceptual Model for Bibliographic Information." The Hague: International Federation of Library Associations and Institutions (IFLA) Functional Requirements for Bibliographic Records (FRBR) Review Group. https://repository.ifla.org/handle/123456789/40.