Annotating Research Infrastructure in Scientific Papers: An NLP-driven Approach

Seyed Amin Tabatabaei, Georgios Cheirmpos, Marius Doornenbal, Alberto Zigoni, Veronique Moore, and **Georgios Tsatsaronis**
Elsevier
s.tabatabaei@elsevier.com

Abstract

In this work, we present a natural language processing (NLP) pipeline for the identification, extraction and linking of Research Infrastructure (RI) used in scientific publications. Links between scientific equipment and publications where the equipment was used can support multiple use cases, such as evaluating the impact of RI investment, and supporting Open Science and research reproducibility. These links can also be used to establish a profile of the RI portfolio of each institution and associate each equipment with scientific output. The system we are describing here is already in production, and has been used to address real business use cases, some of which we discuss in this paper. The computational pipeline at the heart of the system comprises both supervised and unsupervised modules to detect the usage of research equipment by processing the full text of the articles. Additionally, we have created a knowledge graph of RI, which is utilized to annotate the articles with metadata. Finally, examples of the business value of the insights made possible by this NLP pipeline are illustrated.

1 Introduction

According to the definition adopted by the European Commission (European Commission et al., 2012), Research Infrastructure (RI) refers to *"facilities, resources and related services that are used by the scientific community to conduct top-level research in their respective fields and foster innovation"*. A similar concept, which is more commonly used in the United States, is *"Research Core"* (Bai and Schonfeld, 2021).[1] RI plays a crucial role in conducting high-quality research, with significant financial resources invested every year; for example, European countries have invested over 10 billion EUR every year in the period $2014 - 2020$ (European Commission et al., 2019), while UK Research and Innovation (UKRI), the main public research funding agency in the United Kingdom, has announced in its *Corporate Plan* for the years $2022 - 2025$ to increase the RI investments by at least £200 million every year, to reach over £1.1 billion in 2024 to 2025 (UKRI, 2022). It is, therefore, extremely important for all stakeholders in the research landscape to assess the impact of such investments. Various frameworks for impact evaluation have been proposed in the past (OECD, 2019; Griniece et al., 2020) and they all include scientific outputs, particularly publications in peer-reviewed journals, as an important facet of impact.

There are several challenges in tracking research outputs enabled by RI, such as the lack of a standard approach to recognize contributions of facility managers and staff scientists (Bai and Schonfeld, 2021), or the fact that sometimes it is not even considered appropriate to include them as co-authors (Hockberger et al., 2018). Another important issue is that the contribution of RI to the research project is mostly found in the full text of publications, usually in sections named *"Materials and Methods"*, *"Experimental Setup"*, or similar. This means that abstract and indexing databases such as *PubMed*, *Scopus* or *Web of Science*, which don't index the full text of records, are of limited help in this scenario. Other approaches, such as assigning persistent identifiers to scientific instruments and reference them in the manuscript (Stocker et al., 2020), while in principle effective for new publications, require widespread adoption among publishers, as well as time and effort to create a database of equipment records. For all these reasons, the identification of links between publications and RI remains largely a manual and inefficient task (Strubczewski, 2019).

In this work we present a solution to the problem of identifying and linking RI in the text of scientific publications, introducing a pipeline designed to connect scientific publications with RI utilized

[1] We will use *"RI"* throughout the text to refer to both concepts.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 457–463

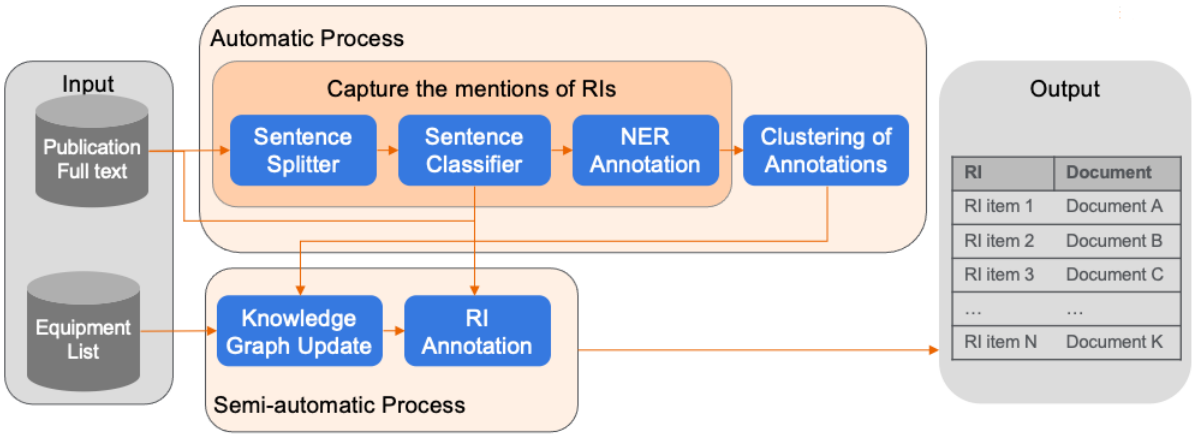


Figure 1: Visual representation of the proposed system. The diagram illustrates the workflow of the pipeline, with each module explained in its respective subsection. Input specifications can be found in section (3.1), and the output of the system is described in (3.2). The various modules include Sentence Splitter (4.1), Sentence Classifier (4.2), Named Entity Recognition (NER) (4.3), Clustering of Annotations (4.4), Knowledge Graph Enrichment (4.5), and RI Annotation (4.6).

in the respective works. To the best of our knowledge, this is the first comprehensive solution to tackle this challenge end-end, and to be brought in a production environment. The pipeline of the solution utilizes state-of-the-art few shot learning algorithms to train our machine learning models using a limited labeled dataset. By employing these cutting-edge techniques, we were able to achieve very insightful results for research stakeholders, despite the constraints of a small training set.

The remaining of the paper is organized as follows; the key user problems that this solution is addressing are discussed in Section 2. Section 3 provides an overview of the solution's architecture, followed by a detailed description of each module in Section 4. Section 5 presents evaluation results and analysis. Section 6 highlights concrete business impact of the proposed system and in Section 7 we conclude and provide pointers to future work.

2 Description of User Problems

Working with academic institutions and funding agencies with a specific interest in RI, we have identified several use cases where being able to link RI (as an input to research) with research outputs (in particular scientific publications) can provide valuable insights to a broad range of stakeholders such as policy makers, academic leaders, researchers and technical staff, as well as the general public. Here are some representative use cases: (1) Supporting decision making processes and investment planning about RI with a quantitative, evidence-based approach that complements qualitative insights based on expert opinion; (2) Showcasing RI to attract top talents. Institutions can promote themselves as a destination for the best researchers in various fields by showcasing state-of-the-art instruments available at their research facilities; (3) Promoting collaboration at local, national and regional level, as well as across disciplines and sectors (for example academic-corporate collaborations); (4) Supporting Open Science and Big Science, by promoting transparency and accountability, particularly for large RIs; (5) Improving the reproducibility of research, by providing useful information about the equipment used in research works.

3 Overview of the Proposed Solution

This section provides an overview of the proposed solution, including the input requirements and desired output of the pipeline. We also discuss the dataset collected for training the NLP components. An overview of the solution's pipeline is illustrated in Figure 1. The first part of the pipeline (unsupervised provess) is composed of fully automated modules, while the latter (supervised process) requires supervision by a *Subject Matter Expert* (*SME*) to guarantee high-quality results.

3.1 Input Specification

The system requires two inputs: (i) the text of academic publications, to identify mentions of RIs,

and, (ii) a knowledge graph of RIs, which is a list of equipment with specified attributes. This list can be customized to a specific research center/university by using their research information management system (e.g., PURE[2]) or equipment/lab management systems (e.g., ClusterMarket[3]).

3.1.1 Full text of publications

The system can handle input text several formats, namely plain-text, XML, and PDF. The section tags of the XML files, when available, can be also utilized for selecting specific sections to process. To transform PDF files into plain-text, the *TIKA* Python library was employed (Apache Software Foundation, 2021).

3.1.2 RI Knowledge Graph

A *Research Infrastructure Knowledge Graph* (KG) is also required, to formalize the representation of participating research institutions, their facilities, equipment vendors and equipments. The equipments facet is organized in broad categories such as *Measuring equipment*, hosting a poly-hierarchy of equipment types, such as *Spectrophotometer*, with equipment models as leaf nodes e.g., *NanoDrop ND*-1000. Each equipment model is linked to: their equipment type(s), the facility and research institutions they are located in, their vendor, the original research institution's local identifier, and, their related method (e.g., *Spectrophotometry* in our previous example).

The KG has been built iteratively and is updated frequently based on customers' needs. After an initialization based on generic lists of equipment types used in the first participating universities from a pilot that was conducted, each customer gives us their actual list of equipment models and types and we place them accordingly in the KG: for each customer, we expand the equipment types hierarchy, using sub-string matching and transformer based clustering methods (using *BERT*) to identify where to automatically place the new RI instances.

3.2 Output Specification

The final output of the pipeline is a table connecting RIs to relevant publications (e.g., DOIs). It's important to note that the relationship between RIs and publications is *many-to-many*, meaning that a single RI can appear in multiple studies and multiple RIs can be used in one study. The resulting table can serve as the foundation for different dashboards, analyses and decision support systems.

3.3 Datasets

The lack of widely available training data for the NLP modules of the pipeline, and the cost of compiling large new data sets for the task has lead us to assemble a small dataset to use in a few shot learning fashion. We used 103 research publications, with 78 being held for training and 25 for model evaluation. To train and test the sentence classifier, all sentences in these publications containing at least one RI were labeled as positive and the rest as negative. However, this resulted in a heavily biased dataset, with less than 3% of the sentences being labeled as positive. The final dataset comprises more than 14K sentences, two-thirds of which were utilized for training the models, while the remaining third was reserved for evaluation purposes. To train the Named Entity Recognition component, we used 354 sentences with annotations provided at the word level. This dataset includes 494 RIs. Using the tokenization process described in subsection 4.1, each word-token was matched to its corresponding label in *BIO* format (Ramshaw and Marcus, 1999).

4 Modules of proposed solution

This section provides the details of each of the components in the pipeline illustrated as blue boxes in Figure 1.

4.1 Sentence Splitter

The initial step in identifying mentions of RIs is to split the full text of a publication into sentences. We use the *Stanza* Python library (Qi et al., 2020) for sentence splitting, as it has very high reported accuracy, but slow processing time. This choice is crucial for correctly identifying RI mentions, as RI names often contain punctuation that could lead to incorrect results with regular expression methods. Additionally, *Stanza* considers not only punctuation, but also contextual meaning, making it more precise, which comes at the cost of slower processing speed compared to other approaches.

4.2 Sentence classifier

The next step is to identify sentences that discuss the usage of RIs using the trained sentence classifier. This step is crucial as not all references to RIs are related to their usage in the current research;

[2] https://www.elsevier.com/solutions/pure
[3] https://clustermarket.com/

for example, authors may compare their work with others'.

For the sentence classification objective a *BERT* for sequence classification, namely SciBERT-base-uncased (Beltagy et al., 2019) pretrained model (Devlin et al., 2018; Wolf et al., 2019) was used under a contrastive loss (CL) objective (Gunel et al., 2020). The sentence classifer attempts to differentiate between samples that not only contain an RI, but also express usage of an RI as context. This improves the overall precision of the model as well as providing valid predictions for the NER module. For the loss, it is known that cross entropy by itself is a weak measurement of loss in a few-shot set-up where labeled data is limited (Dodge et al., 2020; Zhang et al., 2020), thus, we used a normalized summation of the *Cross Entropy loss* and *Supervised Contrastive loss*. Contrastive loss is a technique used in few-shot learning to train models by maximizing similarity between the representations of samples from the same class and minimizing similarity between the representations of samples from different classes. The model was trained for 20 epochs with a batch size of 64, on a $70:30$ split. Learning rate was $1e-5$. Input tokenized vector size per sentence was of maximum length 128. The contrastive loss setting temperature was set to 0.3 and the λ parameter to 0.9.

4.3 NER

Once the sentences discussing the usage of RIs in the research have been identified, a NER component is employed to extract of the RI entities within these sentences. For the entity detection objective a *BERT* for token classification, namely bert-base-uncased pretrained model with a similar contrastive loss objective per above, was used. In this case, the contrast is introduced on the word-token level of the sentence. While the data are scientific publications, parts of the name of an RI can also be found in the common language and there is no base rule for referencing it. It is important to identify, based on the context, which token belongs to an RI and group them together. The larger and more diverse training corpus of BERT-base makes it more sensitive to a broader range of linguistic pattern and contexts over SciBERT which is exclusively trained on computer science and biomedical publications. Empirically, BERT-base was a better candidate for fine-tuning on the downstream NER task due to the representation capturing general language knowledge, despite SciBERT outperforming it in various benchmarks.

As for the word level tokens, the *B* token in this case not only assists in not confusing the individual RIs that were found but also visualizes the model's behaviour on the boundaries it identifies between tokens surrounding the RI, which is done with the assistance of contrastive loss. This model was trained for 28 epochs with a batch size of 8. Input tokenized vector size per sentence was of maximum length 128. Learning rate was $5e-5$.The contrastive loss setting temperature was set to 0.5 and the λ parameter to 0.8.

4.4 Clustering of Annotations

The application of the sentence classifier and NER modules on the input documents results in a large number of mentions of RIs. Due to the variety by which authors cite or quote the equipment used, some of these mentions may match the official names of RIs in a provided supplied equipment list, while others may not. To accurately match these alternative names to a specific RI, we apply a clustering algorithm to group them together. A three-step divide and conquer strategy has been developed to guarantee the correctness of the clustering process. By doing this, we can make sure that references of RIs with different vendor names or model numbers do not fall into the same cluster. This approach is as follows:

1. Group all mentions based on the vendor name mentioned in them. There is also a separate group for mentions without a vendor name.

2. Group the items within each group from step 1 based on the longest word token that contains at least one digit. This substring usually represents the equipment's model number.

3. Each group from step 2 is clustered using *K-means* clustering on the *TF-IDF* representation of the mentions. To find the optimal number of clusters in each group, the *Silhouette* score is maximized.

An example of resulted clusters is presented in table 1

4.5 Knowledge Graph Enrichment

The equipment list from a given university, mapped to our KG's unique identifiers, is used to identify mentions of these pieces of equipment in research

mentions of RI
transmission electron microscope (TEM, JEOL JEM-2010)
JEM-2010 microscope (JEOL, Japan)
JEOL JEM-2010 electron
JEOL JEM-2010 electron microscope
High-resolution transmission electron microscopy (HR-TEM) system (JEOL, JEM-2010)
JEOL JEM-2010
Selected area electron diffraction (SAED, JEOL. JEM-2010, 200.0 KV)
JEOL JEM-2010 transmission electron microscope

Table 1: An example cluster of mentions of RI.

Accuracy	Precision	Recall	F1
0.99227	0.86734	0.7798	**0.82125**

Table 2: Performance of the Sentence classifier model.

articles. These mentions are clustered as described in 4.4. Each cluster is carefully reviewed by a Subject Matter Expert and possibly edited before being added to the KG as synonyms for the equipment data point it is mapped to. Our latest KG version contains over $1,500$ pieces of equipment (types and models) and over $2,500$ vendors.

4.6 Annotation

For extracting the list of RI mentions in a given document set we combine three sources of equipment names into a single vocabulary for text matching: (i) reference vocabulary, cf. 3.1.2; (ii) terms found in the input texts, cf. 4.4; (iii) user-submitted list of equipment of interest. Research institutions or funders have an interest in tracking the usage of equipment that they own or manage and will submit a list of RI that reflects that interest. This list will be matched against the enriched vocabulary, resulting in a final reference list of RI, containing many name variants, and formatted in a way suitable for use in the annotation tool; we employ the annotation tool FPS (Fingerprint Services) that is described by Kohlhof et al. (2014). Applying the RI annotation as a final stage to the process accomplishes several things: (1) it integrates the knowledge accumulated in previous stages; (2) drawing on the FPS capabilities, it allows us to influence recall and precision; (3) it results in a list of consumer-relevant data linked to the right identifiers; (4) applying to specified parts of the full-text documents we can evaluate the quality of the annotation tool for different scenarios, i.e., when applied to "positive" sentences only, when applied to specific text sections only (as explained in 3.1.1 having the input publication in XML format enables us to focus on specific sections, like *Material and Methods*), or when applied to whole text for maximum recall.

Tag	Precision	Recall	F1
O	0.88	0.88	0.88
B-EQ	0.93	0.94	0.93
I-EQ	0.98	0.97	0.98
Macro average	**0.93**	**0.93**	**0.93**
EQ (phrase level)	0.76	0.77	0.77

Table 3: Performance of NER model at tag level and phrase level.

5 Performance Evaluation

5.1 Inference Time

Tested in a sample of 120k full text scientific publications, the total inference time for the complete pipeline, by means of aggregating the inference times for the sub-modules of sentence splitting, sentence classification, NER and clustering of annotations, results to 35.5 hours in a *g4dn.2xlarge* Amazon EC2 instance. The majority of the inference, amounting to 83% is taken up by the first two modules, while NER needed 30 minutes (1.4%) of total elapsed time to complete the processing of all documents.

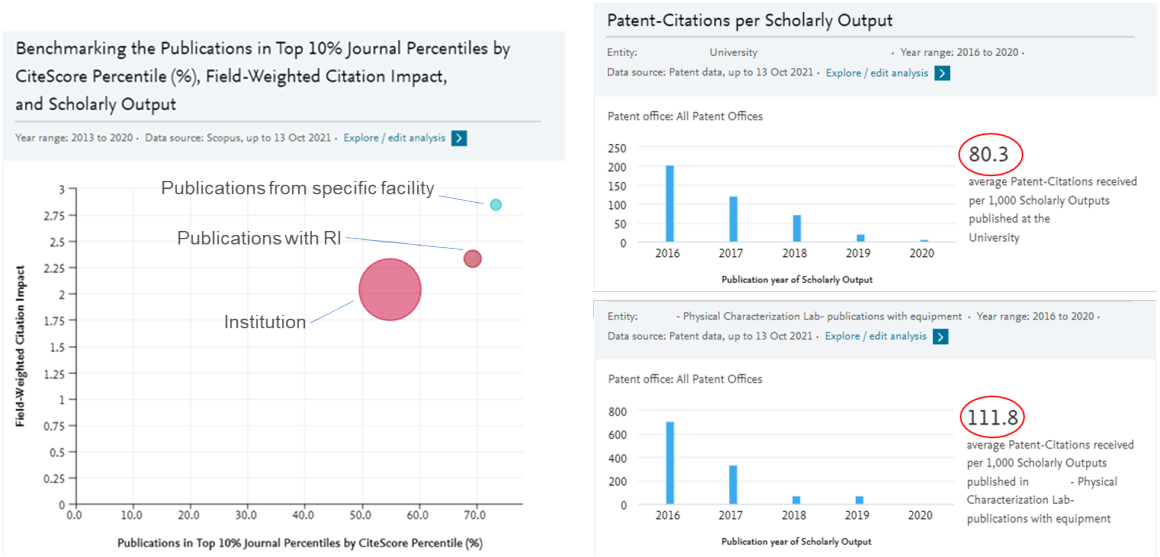
5.2 Precision and Recall of Modules

5.2.1 Sentence classifier

The combination of a scientific BERT model with the contrastive loss assists the sentence classification model to capture the context of RIs utilization. In production state this model parses millions of sentences averaging similar metrics. In Table 2 we present the overall performance of model, as this was measured on our hold-out test set.

5.2.2 NER

Despite the low number of training samples, the NER model with its contrastive nature is able to generalize with very satisfactory performance. The performance in the tags of interest is high enough so that the full extraction of a RIs can be done with a simple post processing of the NER's output. The performance of the NER module in our hold-out test set is shown in Table 3.



(a) (b)

Figure 2: (a) Citation impact of publications from the entire institution, compared to that of publications with associated RI and publications enabled by RI in a specific facility; (b) Average number of patent citations to scientific publications for every 1000 publications, for the entire institution and for publications involving a specific facility.

5.2.3 Discussion on the Overall performance

The performance of the individual models has still room for improvement. In the two previous sections we presented the performance of two of the key components, which is the sentence classifier and the NER. The majority of the issues we observed in a manual error analysis results from the poor generalization of the models in capturing all possible ways of how a RI is reported and discussed in a scientific paper. The sentence classifier component's nature is to lighten the processing load on the NER component. i.e., instead of processing all sentences of a scientific publication for RI entities, to only focus on the ones that the sentence classifier believes they discuss the usage of RI. This additional step also introduces some errors; however, even with an imperfect sentence classifier the NER is still able to distinguish the proper mentions of RI as seen by the high scores in Table 3. Taking into account the phrase level score, it should be highlighted that the NER task is more difficult compared to the conventional NER tasks with common entities like PER/ORG/LOC. In an industrial setting, we have found that the aforedescribed performance is already sufficient to address business use cases and generate very meaningful insights for the RI stakeholders, examples of which we share in the next section.

6 Business Impact

We have completed several projects with institutions active in the Science, Technology, Engineering and Mathematics (*STEM*) domains. We focused primarily on the first use case of those listed in section 2: we helped institutions evaluate the impact of their RI investments by providing quantitative evidence based on a scientometric analysis of publications enabled by institutional RI. Those insights include: (1) the contribution of RI to the scientific output in a certain topic; (2) the scientific impact of publications enabled by RI, compared to the institutional average; (3) the scientific impact of a specific facility or lab inside the institution; (4) the impact on innovation that is enabled by a certain technology available at the institution, and, (5) the role of institutional RI on collaborations with corporate entities. Charts in Fig. 2, which are taken from a report that was done for one of the pilot institutions, illustrate how these insights can be derived using our system.

The evaluation of scientific impact is routinely done by analysing citation networks, and several metrics have been developed for this purpose (Waltman, 2016). The chart in Fig. 2a compares for a specific facility the citation impact of publications from the entire institution, with the subset of publications linked to RI and with a subset of publications linked to equipment. The X axis measures the citation impact of the journals hosting the

publications; the Y axis reports the direct citation impact of the publications. Both metrics are size-independent and normalized. Figure 2b shows that RI is a net contributor to the scientific impact of the institution, as captured by both metrics, as well as how research enabled by equipment from a specific facility has a much higher ratio of patent citations than the institutional average.

7 Conclusions and Future Work

In this paper we have presented a novel system that can detect, extract and link the Research Infrastructure (RI) used and mentioned in scientific publications. The system comprises several advanced NLP components that can annotate and classify sentences, as well as detect RI entities and link them to a knowledge graph (KG) that has been created for the purpose of this business application. We have discussed the performance of the key individual components of our system, the use cases that the proposed solution can address, and we have demonstrated the insights and knowledge that any research facility or institution can obtain around the impact and Return of Investment of their equipment in research conducted by its personnel. Our future work will focus on expanding and releasing the KG in public, as well as optimizing the paralelization and scaling of the existing pipeline.

References

Apache Software Foundation. 2021. Apache tika. https://tika.apache.org/. [Software].

Yuzhou Bai and Roger Schonfeld. 2021. What is a research core? a primer on a critical component of the research enterprise.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

European Commission, Directorate-General for Research, and Innovation. 2012. *Developing world-class research infrastructures for the European Research Area (ERA) : report of the ERA Expert Group*. Publications Office.

European Commission, Directorate-General for Research, and Innovation. 2019. *Research infrastructures make science happen*. Publications Office.

Elena Griniece, Jelena Angelis, Alasdair Reid, Silvia Vignetti, Jessica Catalano, Ana Helman, Matias Barberis Rami, and Henning Kroll. 2020. Guidebook for Socio-Economic Impact Assessment of Research Infrastructures.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Philip Hockberger, Jeffrey Weiss, Aaron Rosen, and Andrew Ott. 2018. Building a sustainable portfolio of core facilities: a case study. *Journal of Biomolecular Techniques : JBT*, 29:79–92.

I. Kohlhof, B. Kozlov, and M. Doornenbal. 2014. Activating qualified thesaurus terms for automatic indexing with taxonomy-based wsd. *Computational Linguistics in the Netherlands Journal*, 4:17–28.

OECD. 2019. Reference framework for assessing the scientific and socio-economic impact of research infrastructures.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. *Natural language processing using very large corpora*, pages 157–176.

Markus Stocker, Louise Darroch, Rolf Krahl, Ted Habermann, Anusuriya Devaraju, Ulrich Schwardmann, Claudio D'Onofrio, and Ingemar Häggström. 2020. Persistent identification of instruments. *Data Science Journal*, 19.

Noelle Strubczewski. 2019. Shared resource facility market analysis.

UKRI. 2022. Ukri corporate plan 2022 to 2025.

Ludo Waltman. 2016. A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2):365–391.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.