Hidden in Plain Sight: Building a Global Sustainable Development Data Catalogue

James Hodson[1] and Andy Spezzatti[2]

[1] AI for Good Foundation, El Cerrito, USA
hodson@ai4good.org
[2] UC Berkeley, Berkeley, USA
andy_spezzatti@berkeley.edu

Abstract. Modern scientific research for Sustainable Development depends on the availability of large amounts of relevant real-world data. However, there are currently no extensive global databases that associate existing data sets with the research domains they cover. We present the *S*DG Data Catalogue, an open, extensible, global database of data sets, metadata, and research networks built automatically by mining millions of published open access academic works. Our system leverages advances in Artificial Intelligence and Natural Language Processing Technologies to extract and organise deep knowledge of data sets available that is otherwise *h*idden in plain sight in the continuous stream of research generated by the scientific community.

Keywords: Artificial Intelligence · Open Data · Information Extraction · Entity Linking · Natural Language Processing

1 Introduction

In 2015, the Unied Nations transitioned from the Millennium Development Goals (MDGs) to the Sustainable Development Goals (SDGs) [2]. The SDGs cover seventeen interconnected goals that define global quantifiable objectives across the social, economic and environmental dimensions of development. They aim to be a universal set of indicators and a reference framework to be leveraged by the global community to motivate policies and implementation by 2030.

This framework supports a long term transition towards more sustainable development. It fosters accountability while also promoting global collaboration. It is a tool to guide decision making but not in and of itself a prescriptive and actionable guide. As a result, member states and institutions are free to pursue policies and programmes according to their own context, resources, and available *s*cientific evidence.

Lack of data is one of the principal bottlenecks in making progress towards the Goals. This problem manifests itself both through (i) the difficulty of consistently measuring *p*rogress; and (ii) the difficulty of selecting appropriate interventions. The first problem is an active area of research, with web-based systems

such as the *S*DG Trendscanner[3] recently maturing to the point of being generally useful. The second problem is not specific to the Sustainable Development Goals, it is an issue that plagues practically all empirical sciences, and has only become more evident as the web has grown.

Researchers do not do a good job of sharing data, which creates barriers to the reproduction of results, and makes it more difficult for scientists to build on each others' work. There are various motivating factors that might arguably explain the current equilibrium: (i) data has become the main innovation in many fields, so researchers seek to publish as many papers as possible on a data set before sharing it, (ii) many data sets are provided through special relationships with commercial entities who need to maintain control of their assets, (iii) departments, universities, states, and countries all may have competitive reasons to keep data private, and (iv) the available mechanisms for sharing data are often not completely open themselves, perhaps even being set up as commercial entities and having burdensome requirements or fees.[4]

This paper's main contribution is focussed on the issue of *d*ata discovery and *d*ata sharing to accelerate progress on the Sustainable Development Goals. Our aim is to provide a system that automatically identifies, aggregates, and describes data sets of relevance to each SDG, performs this task at a global scale, and provides ways to help researchers search for and obtain data sets for their work on Sustainable Development topics.

We leverage the intuition that most published research must provide sufficient details of the data sets used in order to pass any level of peer review. We do not expect authors of empirical work to do anything more than they already do to describe their data sets in their own papers, and publish their papers. It is the system's responsibility to read scientific publications, identify the data sets used, understand the data set in context, and decide whether or not it is relevant to a particular Sustainable Development Goal. Furthermore, the system must be able to associate details of a particular data set across multiple publications, and must be able to identify the owners and contact details of responsible parties where available. A limited number of data sets are truly open, in which case future researchers would be able to download them rather easily. More than 95% of the data sets identified by our system have no further information available beyond what is written in the published work, despite being highly relevant to solving some of the world's most pressing challenges. In these cases, it is our hope that the data owners will be open to sharing their data when a peer scientist places a request relating to the Sustainable Development Goals, in return for appropriate citations.

[3]The SDG Trendscanner, developed collaboratively by the United Nations Development Programme (UNDP), the RISE Research Institutes of Sweden, and the AI for Good Foundation, leverages Artificial Intelligence to aggregate as much global information as possible regarding progress towards each SDG and a variety of related trends. The system is available freely at https://sdg.trendscanner.online.

[4]See [9] for a more in-depth treatment of this topic. However, 'Open Data' services provided by academic publishers unfortunately often result in the public sharing of data tables from papers, rather than the underlying raw data used in analysis.

The *S*DG Data Catalogue leverages Artificial Intelligence and Natural Language Processing techniques in order to "read" millions of academic papers indexed through the Open Academic Graph (OAG) project (see [10] and [8]). The system retrieves full papers, identifies anchor mentions of data sets, and leverages sections of the document to identify further pertinent details, including the number of samples, ownership, and relevant attributes. Each paper is also classified as related or not to each of the Sustainable Development Goals, allowing data sets to be organised under each main category. Where the same data set is leveraged across multiple publications, the system is able to merge extracted information to retain consistent and cross-referenced records. Data set details can then be presented to users in a query-able and actionable format. We perform evaluation at each stage of the system's component pipeline, and find promising performance at each stage, with overall precision of the system's extractions nearing 89%, and recall being around 72% for data set discovery.

The remainder of this paper proceeds as follows: Section 2 describes the data set used, how it was obtained, and how it is cleaned/filtered in preparation for information extraction; section 3 provides details of our annotation and information extraction pipelines; section 4 describes the system architecture and usage; and section 5 concludes.

2 Data

Our aim is to allow our algorithms to read as many papers as possible from the web and partner data sources. The more documents can be collected, analysed, and extracted, the more confident the system becomes in each individual data set extraction. The foundation for our approach is a flexible web-scraper that is both scalable, light-weight, and copyright–aware. Scalable such that we can keep up with the thousands of new documents available each day, light-weight so that we can deploy additional resources quickly, and copyright–aware so that we do not inadvertently retrieve materials subject to explicit policies that prohibit the use of web spiders or machine reading.

The basis for our web scraper is the extensive database of academic publications and references provided by the Open Academic Graph (OAG) project ([10, 8]). The OAG data set contains information on approximately 240M unique global publications, spanning several decades, and across all domains. Each entry contains title, abstract, authors, citations, year of publication, and venue fields, among others[5]. Not all fields contain valid data, but the coverage is very high (more than 95%) for the core entries we are interested in.

For each paper entry in OAG, we construct a search engine query containing the title and author last names. Query results are filtered to PDF files, and the top 5 results become part of a candidate list. We then proceed to attempting the download of each document in turn until we either (a) successfully retrieve a parse-able PDF file, or (b) reach the end of the candidate list. We leverage a

[5] A URL field is also often present, but tends to link to paywall content such as the IEEE and ACM digital libraries. We therefore do not use these links.

proxy service[6] in order to parallelise our download queues up to 100 times and make maximal use of our servers' available bandwidth. With this approach, we retrieve valid PDF documents 65% of the time, and are able to download 10k valid documents per hour. For the purpose of this paper, we work with a subset of 5M valid PDF documents.

3 Parsing and Information Extraction

There is a large literature spanning several decades covering structured information extraction from textual documents. Previous work deals with (a) retrieving text in the correct intended order and format (e.g. [1], (b) understanding the sections and relational structure of the document (e.g. [5]), and (c) extracting complex information such as tables, image captions, and citations (e.g. [3]).

Our present work falls somewhere across all of the above areas: we need to accurately extract the actual text content of the entire document; we then focus on the identification of the sections of a document that deal with our topic of interest (data sets); we employ specific entity recognition algorithms to extract complex relational information of interest, and aggregate this information across a large number of relevant documents.

Fortunately, for some of these areas there are existing implementations of sufficient quality that we can use off-the-shelf. For example, extracting text from PDF files is handled transparently for us by the Apache Tika Content Analysis and Extraction Toolkit[7]. We find that the validity of extracted text is sufficient (i.e. non-garbled and following the correct flow patterns of the original document) in 83% of cases. Therefore, it is important to note that our document processing pipeline is able to correctly process 54 documents out of every 100, before any advanced information extraction can take place (see Table 1).

Table 1. Documents remaining in the pipeline at each stage of processing

	Documents In	Documents Out	Proportion
Web Scraping	8,000,000	5,200,467	65%
Text Extraction	5,200,467	4,316,652	83%
Data Set Identification	4,316,652	2,706,109	62%

3.1 Data Set Mentions

The first part of our information extraction pipeline isolates the segments of documents that contain some information about data sets. We take a bootstrapping approach to training a text classification model based on Convolutional

[6] ScraperAPI Inc., scraperapi.com

[7] https://tika.apache.org

Neural Networks implemented within the SpaCy NLP Library[8] as depicted in Fig. 1.

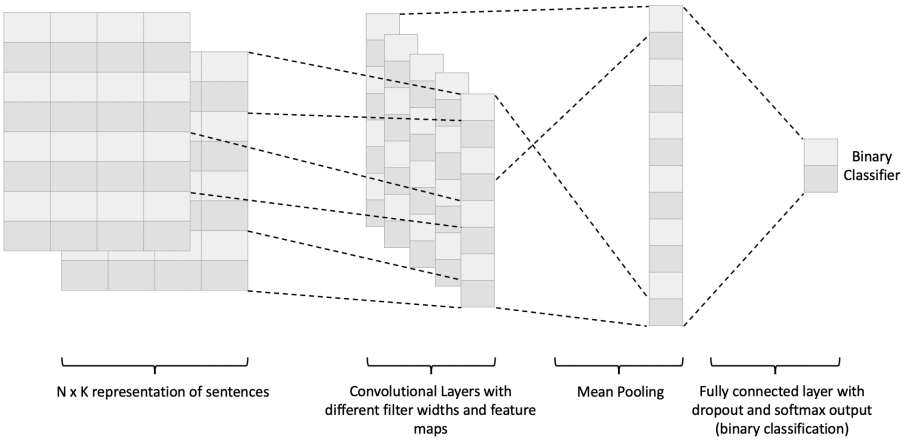

Fig. 1. Text Classification Neural Architecture

First, we extract all paragraphs containing the word 'data', in order to limit the number of candidates for supervised annotation. Next, we identify a series of linguistic patterns as seed terms for prioritising the discovery and annotation of positive examples of document chunks that are relevant to our objective.

Patterns of the form:

```
{the} {<ASCII>*} {data} {set}
{the} {<ASCII>*} {dataset}
...
```

which are highly likely to yield positive examples of data set mentions in the document text. We label these pattern–based examples manually, and use an initial set of 500 annotations to train the text classification model over 10 epochs with a drop-out rate of 0.2. This yields a model with 10-fold cross-validation F-1 score of 74.9%. We then leverage this model to suggest examples for annotation where the uncertainty is highest (where the model's confidence score among the two classes is lowest). We iteratively re-train our model using 300-dimensional word-embedding features trained on the CommonCrawl web-scale data set[9] with the GLobal VEctors for Word Representation (GloVe) procedure (see [7]). Table 2 shows the results when testing against a held-out data set of 10k documents manually annotated for mentions.

[8]See: Honnibal, M, Montani, I. SpaCy NLP Library, v2.3, June 2020. Online: https://github.com/explosion/spaCy

[9]https://commoncrawl.org

Table 2. Identifying Data Set Mentions: We are primarily interested in maximising recall at this stage

	Total Documents	# Data Set Mentions	Recall	Precision	F-1
Data Set Chunks CNN	10,000	5,750	93.5%	48.6%	64%

3.2 Named Entity Recognition

Now that we are able to reliably identify passages of text that contain details of data sets, we move to the next task: extracting specific identifying information about the data. We are interested in identifying the following fields when they appear:

- Data Set Name;
- Data Set Description;
- Owner;
- Number of samples;
- Attribute references;

In addition, we extract document metadata from Tika's Pipeline and OAG database which includes authors, institutions, year of publication, keywords, and publication venue. We also leverage a regular expression to match email addresses in the documents themselves in order to facilitate contacting those with access to or control over the data set.

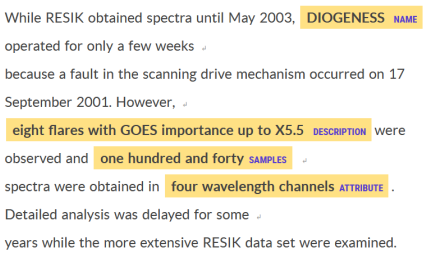

Fig. 2. An example NER annotation for several labels

Each entity of interest from the list above is manually identified through a concerted annotation exercise. We collect a minimum of 500 annotations for each entity within an active learning *m*odel-in-the-loop annotation pipeline (see Fig. 2).

Our Named Entity Recognition model is trained in the same way for each entity, leveraging a Stack-Long-Short-Term-Memory (sLSTM) approach following [6]. This transition–based approach is a relatively recent development in the field of NER, taking its inspiration from the literature on parsing. The results

for each label type are presented in Table 3, and represent testing on a 20% held-out portion of each data set.

Table 3. Performance of NER models, rounded to closest integer value

	Instances	Recall	Precision	F1 Score
NAME	2,725	72%	89%	80%
DESCRIPTION	500	56%	38%	45%
OWNER	500	72%	44%	55%
SAMPLES	500	32%	35%	33%
ATTRIBUTE	742	12%	24%	16%

Performance varies greatly by label category, with worse performance observed for labels that are more heterogeneous. For instance, data set attributes vary widely in type and description, and mentions of data set sample size tend to be very similar to other numeric information that may appear nearby.

Fortunately, the system need not be too sensitive to one-shot extractions. Any data set that has been leveraged in multiple publications may build confidence in its knowledge of that data set through consistent extractions, which are merged.

4 A Sustainable Development Data Catalogue

While this paper represents experiments in building the pipeline for data set extraction to support a global data catalogue, it is worth dedicating some space to motivating and describing the ultimate goal of our work.

A pilot system that leverages text classification models from prior work[4] allows us to accurately associate each academic paper with zero or more of the Sustainable Development Goals. Being able to place each of the extracted data sets into a category, complete with the related papers, author and institution names (and more) provides a data-set discovery system that has the potential to accelerate the application of Artificial Intelligence research on the Sustainable Development Goals. Of course, we wish to take this work further, building query and visualisation layers on top of the *S*DG Data Catalogue (as well as improving all aspects of the document processing pipeline), and allowing researchers to request access to and update details of data sets in the system with minimal overhead. In the end, we hop to create an easily navigable Wikipedia–like experience for keeping track of useful data.

5 Conclusion

The Coronavirus pandemics of 2019 and 2020 have highlighted how the global research community is able to quickly respond to crises and generate new knowledge to help identify solutions. However, the pandemics also highlighted how the

creation of thousands of new papers each day on a single topic can cause more confusion than it might clarify.

Without appropriate intelligent systems it is not possible to keep up to date on new findings or know what data sets to use. The most visible Coronavirus–related data set was a compilation of research papers mentioning "coronavirus", followed closely by rudimentary aggregate statistics from the World Health Organisation (WHO). Clearly, we need to do better to provide our scientific community with the data and tools they need to support decision-making in challenging situations. The Sustainable Development Goals represent a large set of ever more challenging situations, and the *S*DG Data Catalogue is one more resource to help bridge the gap between a set of aspirational objectives, and a clear trajectory.

This paper describes a system that is under active development. The SDG Data Catalogue algorithmic infrastructure will continue collecting new research, improving its coverage and accuracy, and we will be adding new ways to aggregate, sort, and search the resulting information. Our ultimate objective is for the platform to offer incentives that shift the behaviour of research communities toward using their methods first and foremost for the benefit of humanity—bringing more visibility to those researchers and organisations that are most open, collaborative, and productive in developing new solutions to the SDGs.

References

1. Edmundson HP (1969) New methods in automatic extracting. Journal of the ACM (JACM) 16.2: 264-285.
2. Griggs D, Stafford-Smith M, Gaffney O, et al. (2013) Sustainable development goals for people and planet. Nature 495: 305–307. https://doi.org/10.1038/495305a
3. Hodson J. (2016) Table Extraction from Text Documents. In: Sammut C, Webb G (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA
4. Hodson J, Novak B. (2020) Trendscanning for the SDGs. AI for Good Foundation Report. https://ai4good.org/trendscanning Accessed July 2nd, 2020.
5. Iwai I, et al (1989) A document layout system using automatic document architecture extraction. In: Bice K, Lewis C (eds) CHI '89: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA: 369–374. https://doi.org/10.1145/67449.67520
6. Lample G, et al (2016) Neural architectures for named entity recognition. ArXiv Preprint: arXiv:1603.01360.
7. Pennington J, Socher R, Manning CD (2014) GloVe: Global Vectors for Word Representation. In: Moschitti A (eds) Empirical Methods in Natural Language Processing '14. Association for Computational Linguistics, NY, USA.
8. Sinha A, Shen Z, Song Y, et al. (2015) An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion), ACM, New York, NY, USA: 243-246.
9. Stuart D, Baynes G, Hrynaszkiewicz I, et al. (2018) Practical challenges for researchers in data sharing. Springer Whitepaper. https://www.springernature.com/gp/open-research/open-data/practical-challenges-white-paper
10. Tang J, Zhang J, Yao L, et al. (2008) ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference: 990-998.