

1 Potential of natural language processing for metadata extraction from

2 environmental scientific publications

3 Guillaume Blanchy¹, Lukas Albrecht², John Koestel^{2,3}, Sarah Garré¹,

4 ¹Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Melle, Belgium

5 ²Agroscope, Reckenholzstrasse 191, 8046 Zürich, Switzerland

6 ³Institute for Soil and Environment, Swedish University of Agricultural Sciences, Box 7014, 75007 Uppsala, Sweden

7

8 *Correspondence to:* Guillaume Blanchy (guillaume.blanchy@ilvo.vlaanderen.be)

9 **Abstract.**

10 Summarizing information from large bodies of scientific literature is an essential but work-intensive task. This is especially
11 true in environmental studies where multiple factors (e.g. soil, climate, vegetation) can contribute to the effects observed.
12 Meta-analyses, studies that quantitatively summarize findings of a large body of literature, rely on manually curated databases
13 built upon primary publications. However, given the increasing amount of literature, this manual work is likely to require more
14 and more effort in the future. Natural language processing (NLP) facilitates this task, but it is not clear yet to which extent the
15 extraction process is reliable or complete. In this work, we explore three NLP techniques that can help support this task: topic
16 modeling, tailored regular expressions and the shortest dependency path method. We apply these techniques in a practical and
17 reproducible workflow on two corpora of documents: the OTIM and the Meta corpus. The OTIM corpus contains the source
18 publications of the entries of the OTIM database of near-saturated hydraulic conductivity from tension-disk infiltrometer
19 measurements (<https://github.com/climasoma/otim-db>). The Meta corpus is constituted of all primary studies from 36 selected
20 meta-analyses on the impact of agricultural practices on sustainable water management in Europe. As a first step of our
21 practical workflow, we identified different topics from the individual source-publications of the Meta corpus using topic
22 modeling.. This enabled us to distinguish well-researched topics (e.g. conventional tillage, cover crops) where meta-analysis
23 would be useful, from neglected topics (e.g. effect of irrigation on soil properties), showing potential knowledge gaps. Then,
24 we used tailored regular expressions to extract coordinates, soil texture, soil type, rainfall, disk diameter and tensions on the
25 OTIM corpus to build a quantitative database. We were able to retrieve the respective information with 56% up to 100% of all
26 relevant information (recall) and with a precision between 83% and 100%. Finally, we extracted relationships between a set
27 of drivers corresponding to different soil management practices or amendments (e.g. ‘biochar’, ‘zero tillage’, ...) and target
28 variables (e.g. ‘soil aggregate’, ‘hydraulic conductivity’, ‘crop yield’,...) from the source-publications’ abstracts of the Meta
29 corpus using the shortest dependency path between them. These relationships were further classified according to positive,
30 negative or absent correlations between the driver and the target variable. This quickly provided an overview of the different

31 driver-variable relationships and their abundance for an entire body of literature. Overall, we found that all three tested NLP
32 techniques were able to support evidence synthesis tasks. While human supervision remains essential, NLP methods have the
33 potential to support automated evidence synthesis which can be continuously updated as new publications become available.

34 **1 Introduction**

35 The effect of agricultural practices on agroecosystems is highly dependent upon other environmental factors such as climate
36 and soil. In this context, summarizing information from scientific literature while extracting relevant environmental variables
37 is important to establish pedo-climatic specific conclusions. This synthesis is essential to provide recommendations for soil
38 management adaptations that are adequate for local conditions, both, today and in the future. Efforts to synthesize context-
39 specific evidence through meta-analysis or reviews currently requires a lot of manual work to extract specific information from
40 papers. This effort scales with the number of publications which makes it more difficult to collate exhaustive meta-databases
41 from the literature. In the meantime, the use of automated methods to analyze unstructured information (like text in a scientific
42 publication) has been developing during recent years and has demonstrated potential to support evidence synthesis (Haddaway
43 et al. 2020). Natural language processing (NLP) is one of them.

44 In their 2019 review on the advances of the technique, Hirschberg and Manning explained that “natural language processing
45 employs computational techniques for the purpose of learning, understanding, and producing human language content.” This
46 definition is quite broad as NLP encompasses several considerably different techniques, like machine translation, information
47 extraction or natural language understanding. Nadkarni et al. (2011) and Hirschberg and Manning (2019) provide a good
48 overview on this field of research and how it originally developed. For applications to scientific publications, Nasar et al.
49 (2018) reviewed different NLP techniques (information extraction, recommender systems, classification and clustering and
50 summarizations). However, an important limitation of supervised NLP techniques is that they require labels that need to be
51 manually produced to train the model. Hence, humans are still needed for evidence synthesis, but can certainly receive great
52 support from existing NLP techniques.

53 NLP methods are most widely used in medical research. The development of electronic health records significantly facilitated
54 the application of automatic methods to extract information. For instance, information extraction techniques were used to
55 identify adverse reactions to drugs, identify patients with certain illnesses which were not discovered yet at the time or link
56 genes with their respective expression (Wang et al. 2018). A specific example is given by Tao et al. (2017) who used word
57 embedding and controlled random fields to extract prescriptions from discharge summaries. Wang et al. (2018) provide an
58 extensive review of the use of NLP for the medical context.

59 The rise of open-source software tools such as NLTK (Loper and Bird 2002) and SpaCy (Honnibal and Montani 2017) together
60 with the increase in digitally available information has fostered the way for NLP applications towards other scientific
61 communities. For example, SpaCy is able to recognize the nature of words in a sentence and their dependence on other words
62 using a combination of rules-based and statistical methods. Building on that, there are open-source software tools that aim at

63 automatically extracting information. A very popular tool is the OpenIE framework of the Stanford group (Angeli et al. 2015)
64 included in the Stanford coreNLP package (Manning et al. 2014). Niklaus et al. (2018) present a review of open-source
65 information extraction codes. All these tools greatly support novel implementations of NLP applications as they reduce the
66 knowledge required for new users to start using NLP techniques.

67 In the context of evidence synthesis, several NLP methods can be useful. Topic modeling can help to identify common themes
68 in a corpus of publications or classify publications by subject. In addition to selecting publications to be reviewed in the
69 evidence synthesis, topic modelling also gives an overview over the number of publications per topic and helps to identify
70 knowledge gaps. Regular expressions search the text for a pattern of predefined numbers and words. They have a high precision
71 but only find what they are designed to find. This means the user already needs a lot of knowledge on the exact words/terms
72 that should be found. They can be augmented by including syntactic information such as the nature (noun, adjective, adverb,
73 ...) and function (verb, subject, ...) of a word. Complemented with dictionaries that contain lists of specific words (e.g. World
74 Reference Base soil groups), it can be a powerful method. More advanced NLP techniques aim at transforming words into
75 numerical representations that can be further processed by numerical machine learning algorithms. For instance, word
76 embeddings are vectors which encode information about a word and its linguistic relationships in the context it is found. They
77 are derived from the corpus of available documents. An advanced machine learning technique that converts text to a numerical
78 representation are transformer networks such as BERT (Koroteev 2021). BERT transformers are trained on specific corpus.
79 For instance bioBERT is tailored to the medical context (Lee et al. 2020).

80 In contrast to the medical context, fewer studies applied NLP methods to soil sciences. Padarian et al. (2020) used topic
81 modeling in their review of the use of machine learning in soil sciences. Furey et al. (2019) presented NLP methods to extract
82 pedological information from soil survey description. Padarian and Fuentes (2019) used word embedding. They were able to
83 establish relationships between soil types and soil or site properties through principal component analysis. For instance,
84 ‘Vertisols’ were associated with ‘cracks’ or ‘Andosols’ with ‘volcanoes’ as their embeddings were similar.

85 The aim of this study is not to demonstrate the latest and most advanced NLP techniques. Rather, it presents practical
86 workflows to apply NLP techniques to scientific publication in soil science to support different evidence synthesis steps: topic
87 classification, knowledge gaps identification and database building. Our study aims to demonstrate the potential and practical
88 limitations of several NLP techniques through examples of evidence synthesis for soil science. In this regard, we put special
89 emphasis on the methodology used and its ability to recover information, rather than analyzing and interpreting the extracted
90 data itself. We redirect the reader to chapters 1-3 in Garré et al. (2022) for detailed interpretation of the evidence synthesis.

91 The objectives of this paper are (1) to demonstrate the potential of natural language processing as for the collection of structured
92 information from scientific publications, (2) to illustrate the ability of topic classification to identify “popular” and less
93 investigated topics and (3) to assess the ability of natural language processing to extract relationships between a given driver
94 (tillage, cover crops, amendment, ...) and soil variables (hydraulic conductivity, aggregate stability, ...) based on publication
95 abstracts.

96 **2 Materials and methods**

97 **2.1 Text corpora**

98 This work used two corpora (sets of texts) which are referred to in the following as the OTIM and the Meta corpus. The OTIM
99 corpus was related to OTIM-DB (<https://doi.org/10.20387/bonares-q9b3-z989>, EJP SOIL - CLIMASOMA 2022. Chapter 4)
100 which is a meta-database extending the one analyzed in Jarvis et al. (2013) and Jorda et al., (2015). OTIM-DB contains
101 information about the near-saturated hydraulic conductivity obtained from tension-disk infiltrometer between 0 and -10 cm
102 tension . OTIM-DB also includes metadata on the soil (texture, bulk density, organic carbon content, WRB classification), 23
103 climatic variables that were assigned based on the coordinates of the measurement locations, among them annual mean
104 temperature and precipitation, methodological setup (disk diameter, method with which infiltration data is converted to
105 hydraulic conductivity, month of measurement) and land management practices (land use, tillage, cover crops, crop rotation,
106 irrigation, compaction). All data in OTIM-DB were manually extracted by researchers from 172 source-publications. The
107 collected data was then cross-checked by another researcher to catch typos and misinterpretations of the published information.
108 The OTIM corpus consisted of the entire texts of the 172 source-publications used in OTIM-DB.

109 In contrast, the Meta corpus contained only abstracts, namely the one of the primary studies included in the meta-analyses by
110 EJP SOIL - CLIMASOMA (2022) Chapter 1 which investigated how soil hydraulic properties are influenced by soil
111 management practices. This Meta corpus contained 1469 publications. By number of publications, it was substantially larger
112 than the OTIM corpus. The information given in the Meta corpus was not available in a meta-database. Therefore, the
113 validation step had to be carried out by manually extracting information from a subset of the abstracts in this corpus. The
114 references for both, the OTIM and the Meta corpus are available on the GitHub repository of this project
115 (<https://github.com/climasoma/nlp>).

116 **2.2 Extracting plain text from the PDF format**

117 For both corpora, all publications were retrieved as PDF files. The software “pdftotext”
118 (<https://www.xpdfreader.com/pdftotext-man.html>) was used to extract the text from these PDFs. The text extraction worked
119 well apart from one exception where the extracted text contained alternating sentences from two different text columns, making
120 it unsuited for NLP. Other two columns publications were correctly extracted. Other methods were tested, such as the use of
121 the Python package PyPDF2 or the use of the framework pdf.js but did not provide better results than pdftotext. The difficulty
122 of this conversion lies in the PDF format itself that locates words in reference to the page and does not preserve information
123 on which words belong to individual sentences or paragraphs. Recovery methods (such in pdf.js or pdftotext) use the distance
124 between words to infer if they belong to the same sentence and detect paragraphs. This makes extracting text from PDF harder
125 for algorithms and is clearly a major drawback of this format. In addition, text boxes and figures can span multiple text columns
126 and make the conversion difficult (e.g. the figure caption intercalated in the middle of the text). This led Ramakrishnan et al.
127 (2012) to develop LA-PDFText, a Layout Aware PDF to text translator designed for scientific publication (which was not used

128 in this study due to time restriction). The addition of a hidden machine-friendly text layer in the PDF itself or the use of the
129 full-text HTML version can possibly alleviate this issue. Another limitation of the PDF format is that tables are encoded as a
130 series of vertical/horizontal lines placed at a given position. When converting the PDF to text, only streams of numbers can be
131 retrieved. Rebuilding the tables based on the regularity of the spacing between these numbers is possible in some cases (e.g.
132 Rastan et al., 2019) but nevertheless understanding what these numbers represent based solely on the headers is for now out
133 of reach of the NLP algorithm. For this issue too, HTML format has an advantage as tables are encoded in the HTML or
134 provided as separate .xlsx, .csv files, hence enabling easier information extraction.

135 However, because online HTML full-texts were not available for all documents (mostly older publications) and PDF remains
136 the most widespread format for exchanging scientific publication, we decided to pursue the analysis with the PDF format.
137 From the extracted full-texts, abstract and references sections were removed and only the body of the text was used to form
138 the documents for each corpus.

139 Several NLP techniques were applied. Table 1 summarizes which techniques were applied to which corpus.

140 **Table 1: Overview of NLP techniques used and corpus considered.**

Technique	Corpus
Topic classification	Meta corpus
Rules-based extraction (regular expression)	OTIM corpus
Co-occurrence of practices	Meta corpus
Knowledge gap identification	Meta corpus
Relationship extraction	Meta corpus

141

142 **2.3 Topic modeling**

143 Topic modeling creates topics from a corpus by comparing the similarity of the words between documents. We extracted all
144 bigrams (two consecutive words) that occurred more than 20 times for each document in the Meta corpus. Each document was
145 then represented by its bigrams. We found that using bigrams instead of single words help to obtain more coherent topics. This
146 is intuitively explained by considering that the bigrams ‘cover crops’, ‘conventional tillage’ are more informative than ‘cover’,
147 ‘crops’, ‘conventional’ and ‘tillage’ alone. Next, we removed all bigrams that appeared in less than 20 documents and more
148 than 50% of all documents to avoid too specific or too common bigrams. Note that bigrams are usually added to monograms.
149 However, in our case, we found that bigrams alone (e.g. ‘conventional tillage’, ‘cover crops’) led to more coherence topics
150 than when combined with monograms (e.g. topics around ‘tillage’, ‘crops’, ‘water’). Topics were then created to be as coherent

as possible using a latent dirichlet algorithm (LDA). A number of different coherence metrics exist (Röder et al. 2015). In this work, we used the LDA implementation of the *gensim* library (v4.1.2) with the Cv coherence metric. This Cv coherence metric is a combination of a normalized pointwise mutual information coherence measure, cosine vector similarity and a boolean sliding window of size 110 words as defined in Röder et al. (2015). The metric ranges from 0 (not coherent at all) to 1 (fully coherent). To define the optimal number of topics to be modeled, we iteratively increased the number of topics from 2 to 20 and looked at the coherence. Based on this optimal number of topics, the composition of each topic was further analyzed using the figures generated by pyLDAvis package (v3.3.1) which is based upon the work of Sievert and Shirley (2014).

2.4 Rules-based extraction

Regular expressions are predefined patterns that can include text, number and symbols. For instance, disk-diameters of tension-disk infiltrometers are extracted from a text using the regular expression search term '`(\d+\.\d)cm\sdiameter`', which will retrieve text passages like '5.4 cm diameter'. In this regular expression, \d denotes a digit, \s a space, \d+ one or more digits and parentheses are used to enclose the group we want to extract. Regular expressions are a widely used rule-based extraction tool in computer science. They have a high precision but their complexity can quickly increase for more complex patterns. Figure 1 provides examples of regular expressions used in our study. It can be observed that regular expressions for geographic coordinates are quite complex as they need to account for different notations such as decimal format (24.534 N) or degree-minute-second format (24°4'23.03" N) in the case of latitudes. In contrast, specific well-defined terms such as World Reference Base (WRB) soil types are more easily retrieved as their wording is always unique in the text. Soil textures are likewise easy to extract but less well-defined as, for example, WRB soil types. Often, terms used to describe the soil texture of an investigated field site are also used to refer to general cases or unrelated field sites in the same text. This makes it more challenging to automatically extract information on the investigated site using regular expressions. To complicate matters, soil textures are not always given in the USDA (United States Department of Agriculture) classification system, which can be regarded as a standard. For the sake of simplicity, we did not attempt to identify the texture classification system but treated all textural information as if they were using the USDA system. When gathering information on tension-disk diameters, attention on the length units needed to be paid as well as whether the radius or the diameter was reported. In these more complex cases, we constructed the regular expression search terms iteratively to extract the greatest amount of information from the available papers. Regular expressions were used to extract latitude, longitude, elevation, soil type, soil texture, annual rainfall, disk size and tensions applied.

Regular expression matching (e.g.)

The site is situated in 51°46'34.9"N 4°49'12.6"E in the Noordwaard Polder in Netherlands at an elevation of 23 m above sea level.

```
> latitude: (([+-]?[1-8]?\d|[+-]?90)([^*o0])\s?(\d{1,2})(.)\s?(\d{1,2}(\.\d+)?))?.?.?\s?(latitude\s)?([NS])
> longitude: (([+-]?180|[+-]?1[0-7]\d|[+-]?[1-9]\d)([^*o0])\s?(\d{1,2})(.)\s?(\d{1,2}(\.\d+)?))?.?.?\s?(longitude\s)?([WEO]))
> elevation: ((\d+)\s?m[a-z\s]+(altitude|elevation)) or ((altitude|elevation)[a-z\s]+(\d+)\s?m)
```

The soil is a Luvisol (WRB) with a sandy loam texture.

```
> soil type: (Luvisol|Cambisol|Regosol|Podzol|...)
> soil texture: (sandy loam|loamy clay|sand|clay|loam|clay loam|...)
```

Annual rainfall precipitation is 850 mm.

```
> annual rainfall:
((?:cumulated|annual|average)[a-z\s]+(?:rainfall|rain|precipitation))(?:[a-z\s]+)?(\d+[.-\ -]\?\d+c)?[a-z\s]+(\d+\.\?,?\d+(?:-|\sand\s|\sto\s)?(?:\d+))\s?(m\s?m|cm)
```

The measurements were collected using a tension-disk infiltrometer with a 4.45 cm radius at tensions of -1, -3, -5 and -7 cm.

```
> disk size: (radius|diameter)[a-z\s]+(\d+\.\?\d+)\s?(cm|mm) or (\d+\.\?\d+)\s?(cm|mm)[a-z\s]+(radius|diameter)
> tensions: ((?:(?:-?\d+),?\s){2,})\s?(mm|cm)
```

Figure 1: Examples of different regular expressions used for information extraction. \d represents a digit, \s a space, . (a dot) an unspecified character, [a-z] all lower case letters and more generally squared brackets are used to denote the list of characters (for e.g. [^*o0] is used to catch the symbol of degree in latitude/longitude). A character can be “present once or absent” (\d|\.\d? will match both integers and decimal numbers), “present at least once” (\d+ will match 7, 73 and 735) or “present a given number of times” (\d{1,2} will match 7 and 73 but not 735). Parentheses are used to segment capturing groups and can also contain boolean operator such as OR denoted by | which is used to catch exact WRB soil name in (Luvisol | Cambisol | ...). Non-capturing parentheses are denoted with (?:) like for the regular expression of tensions. The content inside non-capturing parentheses will not be outputted as results of the regular expression in contrast to other parenthesis groups.

178

179 To assess the quality of the extraction, different metrics were used. They are illustrated in Fig. 2. For rules-based extraction,
180 two tasks are required by the algorithm: selection and matching. The selection task aims to assess the ability of the algorithm
181 to extract relevant information from the text. The matching task assesses the ability of the algorithm to extract not only the
182 relevant, but also the correct specific information as recorded in the database used for validation. For instance, if the NLP
183 algorithm identified “Cambisol” as the soil group for a study conducted on a Cambisol, both, the selection was true positive
184 (TP) and the matching was true. If the text did not contain any WRB soil type and the NLP did not return any, both selection
185 and matching performed well with the selection being a true negative (TN) case. Eventually, when the NLP algorithm did not
186 find a WRB soil type, but the database listed one, the selection was referred to as false negative (FN) and the matching as false.
187 The opposite case, with a soil type found in the text but no entry in the database was called false positive (FP) and the matching
188 was equally false. Eventually, there were cases where the NLP algorithm retrieved incorrect information but still provided a

189 meaningful value, e.g. if the algorithm extracted ‘Luvisol’ as the soil type while the correct value was ‘Cambisol’. Then the
 190 selection task was still successful since the found term represents a WRB soil type. However, the matching task failed. Such
 191 cases were still marked as true positives, but with a false matching.

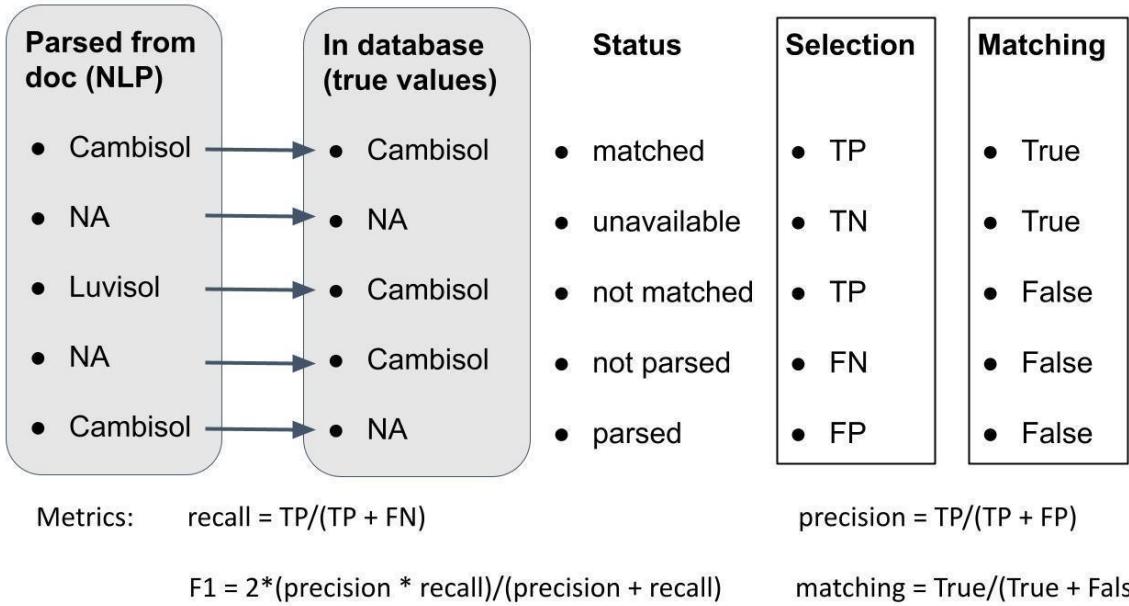


Figure 2: Cases of NLP extraction results in regards to the value entered in the database (considered the correct values) for the selection task and the matching task. TP, TN, FN, FP stand for true positive, true negative, false negative and false positive, respectively. The recall, precision, F1 score and matching values served as metrics for each task.

192
 193 Four different metrics were used to evaluate the results: the recall, the precision, the F1 score and the matching score. The
 194 recall assesses the ability of the algorithm to find all relevant words in the corpus ($\text{recall} = 1$). The precision assesses the ability
 195 of the algorithm to only select relevant words ($\text{precision} = 1$). If there were 100 soil types to be found in the corpus and the
 196 algorithm retrieved 80 words of which 40 were actually soil types, the recall was $40/100 = 0.4$ and the precision was $40/80 =$
 197 0.5. The F1 score combines the recall and precision in one metric which is equal to 1 if both recall and precision were equal
 198 to 1. The recall, precision and F1 scores were used to assess the ability of the algorithm to extract relevant information from
 199 the text. Figure 2 also includes the equations for recall, precision, F1 score and matching score. Note the difference between
 200 precision and matching score: the precision expresses how many relevant words were extracted while the matching score
 quantifies the fraction of words corresponding to the correct information. Considering the example above, if out of the 40

202 correctly selected soil types, only 20 actually matched what was labeled in the document, then the matching score would be
203 $20/100 = 0.2$. Figure A1 gives a graphical overview of the recall and precision metrics. In addition to these metrics, a matching
204 score was used to illustrate how many NLP extracted values actually matched the one manually entered in the database. All
205 rules-based extraction were applied on the OTIM corpus and the information stored in OTIM-DB was used for validation.

206

207 In addition to the above extraction rules, we also identified agricultural practices mentioned in the publications and the co-
208 occurrence of pairs of practices within the same publications. This enabled us to highlight which practices are often associated.
209 To identify management practices in the OTIM corpus, we used the list of keywords from the Bonares Knowledge Library
210 (<https://klibrary.bonares.de/soildoc/soil-doc-search>). Given that several keywords can relate to the same practice, the list was
211 further expanded by using the synonyms available from the FAO thesaurus AGROVOC (Caracciolo et al. 2013).

212 **2.5 Relationship extractions**

213 Relationship extraction relates drivers (agricultural practices) defined by specific key terms to specific variables (soil and site
214 properties). In this study, examples for drivers were 'tillage', 'cover crop', or 'irrigation'. Among the investigated variables
215 were 'hydraulic conductivity', 'water retention' or 'yield'. A list of all considered drivers and variables is given in Table A2.
216 These keywords corresponded to the keywords used in early stages of assembling the Meta corpus (EJP SOIL - CLIMASOMA
217 2022, Chapter 1). To allow catching both plural and singular form, all drivers and keywords were converted to their meaningful
218 root: their lemma (e.g. the lemma of 'residues' is 'residue').

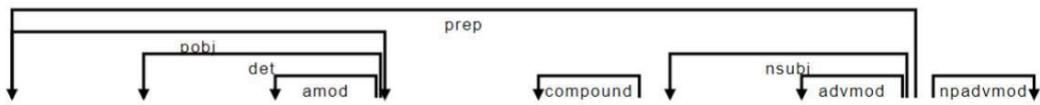
219

220 The relationship extraction algorithm searched in the Meta corpus for sentences which contained lemmas of both, drivers and
221 variables. Each sentence was then splitted in words (tokenized) and each word (token) was assigned a part-of-speech (POS)
222 tag (e.g. noun, verb, adjective). Dependencies between the tokens were also computed. Using these dependencies as links, a
223 graph with one node per token was built. The nodes corresponding to the driver and variables were identified and the shortest
224 dependency path between them was computed (Fig. 3).

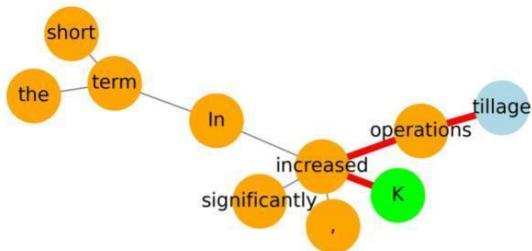
225

In the short term, tillage operations significantly increased K [...].

(a)



(b)



(c)

> verb in common classification:
 positive: 'enhance', 'increase', 'improve',
 'mitigate', ...
 negative: 'decrease', 'reduce', 'prevent',
 'be lower', ...
 neutral: 'have no effect', 'have no significant effect',
 'have minimal effect', ...

Figure 3: Example of NLP extraction on a sentence. (a) shows the part-of-speech (POS) tag below each token and the dependencies (arrows) to other tokens. (b) Based on these dependencies a network graph was created and the shortest dependency path between the driver (blue circle) and the variable (green circle) is shown in red. (c) The verb contained in the shortest dependency path was classified into positive, negative or neutral according to pre-established lists.

226

227 All tokens that were part of this shortest dependency path between the driver token and the variable token were kept in a list.
 228 From this list, the tokens containing the driver/variables were replaced by the noun chunk (=groups of nouns and adjectives
 229 around the token) as important information can be contained in this chunk. For instance the driver token "tillage" was replaced
 230 by its noun chunk "conventional tillage". The list of tokens that constituted the shortest dependency path always included the
 231 main verb linking the driver and the variable token. This verb depicted a positive, negative or neutral correlation between the
 232 driver and the variable. Other modifiers such as negation marks or other modifiers that can be part of the noun chunk (e.g.
 233 'conservation' or 'conventional' with the noun 'tillage') were also searched for in each sentence. In cases where a positive
 234 correlation was negated (e.g. "did not increase", "did not have significant effect on"), the relationship was classified as neutral.
 235 Sometimes, the relationship did not relate directly to the correlation between the driver and the variable but rather mention that
 236 this relationship was studied in the manuscript. Then, the status of the relationships was set to "study". To assess the recall and
 237 precision of the technique, a subset of 129 extracted relationships was manually labeled. Table 2 offers examples of
 238 relationships classified by the algorithm.

239 **Table 2: Examples of relationships identified and their corresponding classified labels.** Note that the modifiers present in the noun
 240 chunk (e.g. "conservation tillage" or "zero tillage") and the negation in the sentence were taken into account in the status of the

241 **relationship.** Some sentences contain multiple driver/variable pairs and, hence, multiple relationships. In such cases, only one of the
242 two was indicated in the table below (but all were considered in the code).

Relationship (driver/variable in bold)	Status
In the short term, tillage operations significantly increased K ($P < 0.05$) for the entire range of pressure head applied [...].	positive
In humid areas, soil compaction might increase the risk of surface runoff and erosion due to decreased rainwater infiltration.	positive
Both tillage treatments were designed to prevent runoff and both increased rainwater penetration of the soil.	negative
After 3 years of continuous tillage treatments, the soil bulk density did not increase.	neutral
No-tillage increased water conducting macropores but did not increase hydraulic conductivity irrespective of slope position.	neutral
A field study was conducted to determine the effect of tillage -residue management on earthworm development of macropore structure and the infiltration properties of a silt loam soil cropped in continuous corn.	study
Dry bulk density , saturated hydraulic conductivity (K_s) and infiltration rate [$K(h)$] were analysed in untrafficked and trafficked areas in each plot.	study

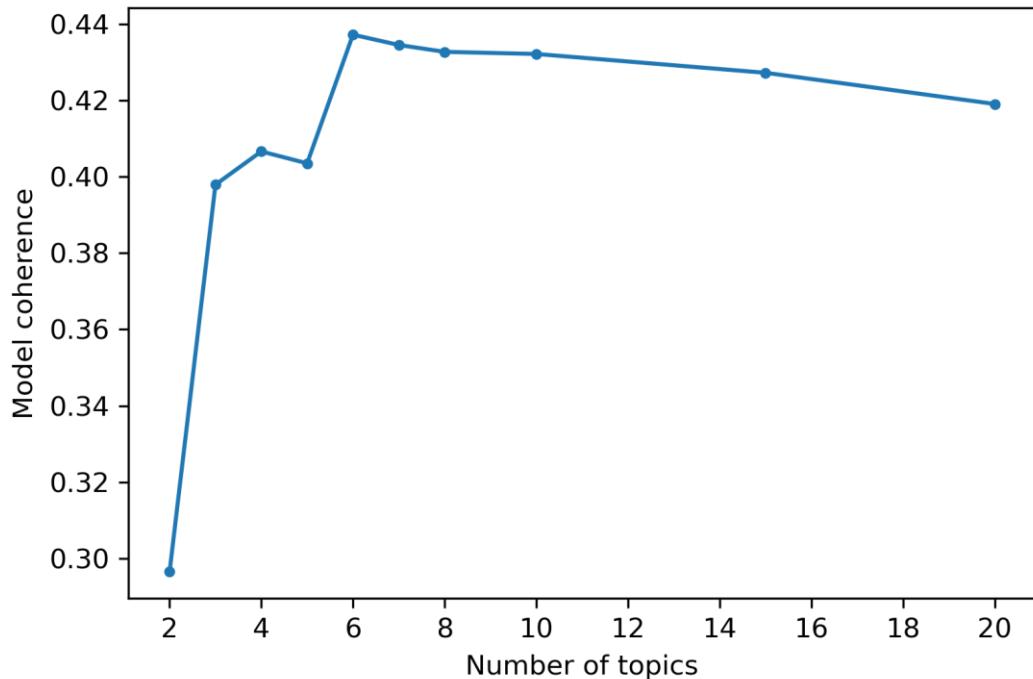
243
244 When identifying driver and variable pairs among abstracts, the case can be encountered where one of the driver/variable is
245 expressed using a pronoun. This prevents keyword-based detection. The *neuralcoref* Python package was used to replace the
246 pronouns by their initial form using co-references. This package uses neural networks to establish a link between the pronoun
247 and the entity it refers to. The pronoun is then replaced by the full text corresponding to the entity. For the Meta corpus, the
248 co-reference substitution did not enable to increase the amount of relevant sentences extracted. It turned out that the use of
249 pronouns in the investigated abstracts was very limited. In addition, the accuracy of the co-reference substitution was not
250 always relevant and substitution errors were more frequent than desired. For these reasons, we left this step out of the final
251 processing pipeline. Nevertheless, we want to stress that replacing pronouns may be very useful for other corpora. Automatic
252 relationships extraction using OpenIE was also tried but given the specificity of the vocabulary in the corpus of abstracts, it
253 yielded relatively poor results.
254 To ensure reproducibility, all codes used in this project were written down in Jupyter notebooks. This enabled the results to be
255 replicated and the code to be reused for other applications. Jupyter notebooks also enable figures and comments to be placed

256 directly inside the document, hence helping the reader to better understand the code snippets. All notebooks used in this work
257 are freely available on GitHub <https://github.com/climasoma/nlp/>.

258 **3 Results and discussion**

259 **3.1 Topic modeling**

260 Figure 4 shows the evolution of the coherence metric with respect to the number of topics. The overall model coherence
261 increases up to 6 topics then slowly starts to decrease. Note that the coherence scores can slightly change between runs as the
262 algorithm starts from a different random seed to build the topics. Nevertheless, we observed a stagnation of the overall
263 coherence after 6 topics, meaning that increasing the number of topics above 6 did not increase the overall coherence of the
264 model.



265
266 **Figure 4: Evolution of overall model coherence according to the number of topics chosen to train the LDA model.**
The coherence metrics is the CV described in Röder et al. (2015) which is a combination of a normalized pointwise
267 mutual information coherence measure, cosine vector similarity and a boolean sliding window of size 110..

265
266 Figure 5 (left) shows the frequency of the topic in the corpus (as percentage of documents in the corpus that belong to this
267 topic). The circles are placed according to the first two principal components based on their inter-topic distance computed

268 using the Jensen–Shannon divergence (Lin, 1991). Topics closer to each other are more similar than topics further apart. Figure
269 5 (right) shows the frequency of each bigram in the topic and in the corpus. Different themes are visible from the topics:
270 microbial biomass and aggregate (topic 1), conventional/conservation tillage (topic 2), crop residue and crop rotation (topic
271 3), water retention and porosity (topic 4), infiltration rate and grazing (topic 5) and cover crops (topic 6).. The left part of
272 Figure 5 shows how topics 1 to 4 are close in contrast to topic 6 that mainly focus on cover crops. These subtopics nicely
273 correspond to some of the main drivers initially set in the search query string used to build the Meta corpus (EJP SOIL -
274 CLIMASOMA 2022, Chapter 1).The topic modeling shows that bigrams such as “cover crops” have a large term frequency
275 (blue bars), which means they are relatively frequent inside the set of documents. Bigrams such as “conservation tillage”,
276 “aggregate stability” or “microbial biomass” are less frequent (smaller blue bars). The topic modeling also shows that terms
277 such as “grain yield” appear in several topics (topic 2 and 3). But it is more frequent in topic 2 than topic 3 (size of the orange
278 bars). On the opposite, bigrams such as “hairy vetch” or “winter cover” are entirely specific to topic 6 on cover crops. It should
279 be also noted that bigrams such as “deficit irrigation”, while present in the corpus, do not appear in the top 6 more relevant
280 terms. This shows that this theme is less represented in the corpus and possibly indicates a knowledge gap around it. Another
281 possible explanation is that, while few papers mentioned “deficit irrigation”, the inclusion of monograms such as “irrigation”
282 might have led to the construction of a topic around irrigation techniques (where papers around ‘deficit irrigation’ might have
283 been found). While different runs of the LDA algorithm led to slightly different topic distributions due the randomness
284 internally used by the algorithm, we observed the appearance of the same coherent topics around tillage, cover crops or biochar.
285 This highlights the fact that the LDA algorithm is not a deterministic method, but a probabilistic one as the probability of a
286 topic per document and the probability of a term per topic are iteratively optimized to maximize the coherence by the LDA
287 algorithm. Overall, we consider that topic modeling can serve as a first tool for an exploratory analysis of the corpus content.

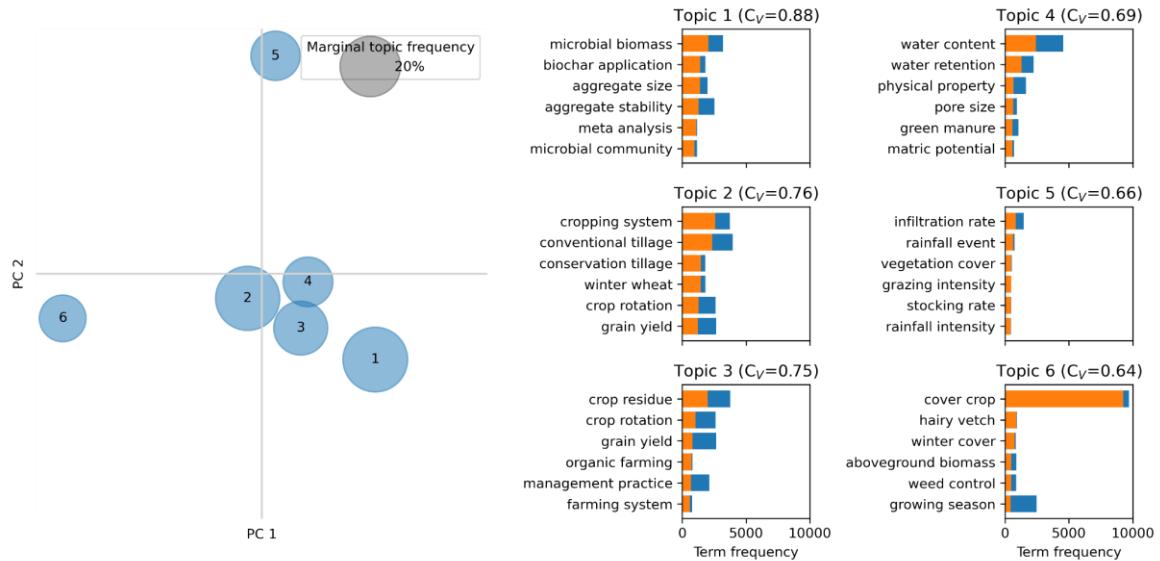


Figure 5: (left) Map of topics according to the first two principal components after dimension reduction. (right) For each topic, the 6 more relevant bigrams inside the topic. The orange bars represent the term frequency inside the topic while the length of the full bars (orange + blue) represent the term frequency in the entire corpus. The gray circle represents the size of a topic that contains 20% of the documents of the corpus.

288

3.2 Rules-based extraction

289

Table 3 shows the metrics relative to the different rules-based extraction techniques. Note that “n” does not always represent 290 the number of documents in the corpus as a document can contain multiple locations for instance. Regular expressions 291 associated with a dictionary for soil texture and soil type provide the best precision overall due to their high specificity. This 292 clearly highlights the usefulness of the international scientific community agreeing on a common vocabulary or classification 293 system. Soil type had the highest recall, which means that all instances of soil types mentioned in the document had been 294 successfully extracted. Regular expression matching quantities such as ‘rainfall’, ‘disk diameter’, ‘tensions’ or ‘coordinates’ 295 had lower recall than rules making use of a dictionary. Coordinates had a high precision but a lower recall as some coordinate 296 format could not be extracted from the text. This could be partly explained by the conversion of the symbols for degree, minute, 297 seconds from PDF to text. As the encoding of these characters varies a lot between journals, the conversion sometimes led to 298 “°” converted to “O”, “*” or “0”. Identifying all these different cases while retaining a high accuracy on more frequent cases 299 was challenging with regular expressions.

300

Table 3: Scores of the rules-based extraction methods. n is the number of items to be extracted. It varies as several coordinates can 301 be provided in the same paper. The method can use only a regular expression (regex) or a combination of regular expression and 302 dictionary (regex + dict.).

Extracted	Method	n	Precision	Recall	F1-score	Matching
Soil type (WRB/USDA)	Regex + dict.	174	0.92	1.00	0.96	0.95
Soil texture (USDA)	Regex + dict.	174	0.95	0.88	0.91	0.83
Rainfall	Regex	174	1.00	0.81	0.90	0.89
Disk diameter	Regex	174	0.83	0.66	0.73	0.41
Tensions	Regex	154	1.00	0.56	0.72	0.31
Coordinates	Regex	209	0.92	0.77	0.84	0.73

303

304 Regular expressions have to be flexible enough to accommodate the various formats found in the publications (e.g. for
 305 coordinates) but also discriminant enough to not match irrelevant items. For instance, the regular expression about soil texture
 306 catches a lot of terms related to soil texture but not all were related to the soil texture of the actual field site. Applying regular
 307 expression on specific parts of the manuscript (for instance, just on the material and methods section), could help improve the
 308 precision of the technique. Note that the regular expression algorithm itself is infallible by nature (it will always return exactly
 309 what is matched). Rather, here, we assess our ability to generate regular expression patterns that are general enough to extract
 310 information for most cases. Adjusting the regular expression to fit all edge cases encountered is, in theory, possible but will be
 311 work intensive and will not scale well with an increasing number of papers.

312 In addition to extracting specific data, general information about which management practices are investigated in the studies
 313 is also important. Figure 6 shows the co-occurrence of the detected practices inside the same document as the percentage of
 314 documents in the OTIM corpus that contains both practices. For instance, the practice of 'crop residue' and 'conversion tillage'
 315 is often found with documents that contain 'conventional tillage'. This can be put in parallel with the topic modeling where
 316 these two bigrams were often associated. 'herbicide' is also often mentioned with documents containing 'crop residue'. Given
 317 the small size of the chosen corpus, the co-occurrences need to be interpreted in connection to experimental sites chosen for
 318 tension-infiltrometer measurements and hence provide an overview of which practices have been most studied with tension-
 disk infiltrometers.

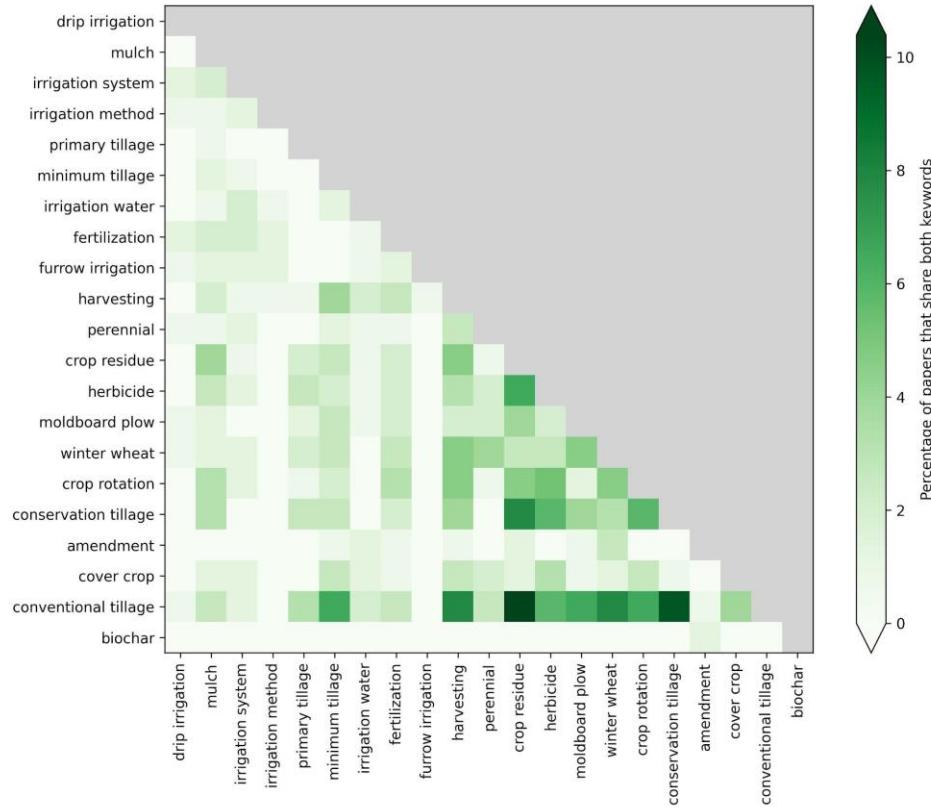


Figure 6: Co-occurrence matrix of identified management practices from the OTIM corpus.

320

321 **3.3 Relationship extraction**

322 Figure 7 shows the number of relationships from abstracts extracted according to the pair driver/variable identified within
 323 them. Relationships including “biochar” or “tillage” as drivers were the most frequent while “yield” was the variable most
 324 commonly found. Note as well that for some combination of drivers/variables, no statements were available. This helped to
 325 identify knowledge gaps within our corpus. For instance, the effect of liming on aggregates and infiltration properties was not
 326 studied in our corpus. Similarly the effect of irrigation on soil organic carbon was also not present in the corpus.

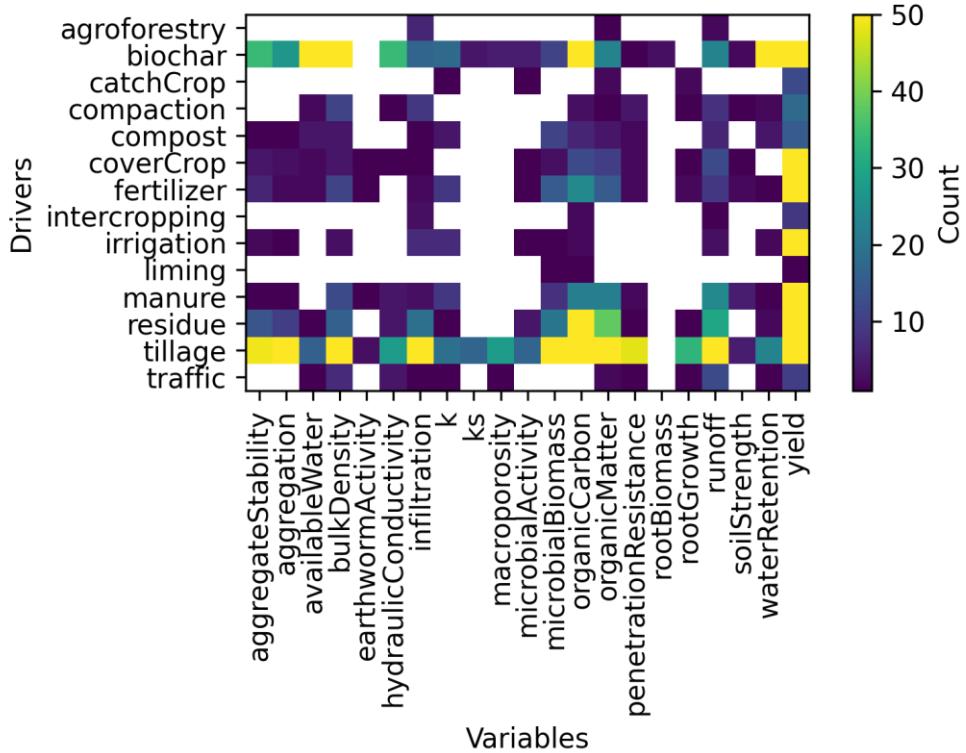


Figure 7: Number of relationships identified from abstract according to the pair driver/variable they contain. White cells mean that no relationships were found for the pair inside. Results obtained from the analysis on the Meta corpus.

327

328 However, one important limitation of the approach is that the algorithm can only find the keywords it was told to look for. For
 329 instance, no social drivers were found in the statements as there were no keywords associated with it. Social drivers are
 330 important to estimate the acceptability of management practices (EJP SOIL - CLIMASOMA 2022, Chapter 3) and they would
 331 gain to be included in the workflow. Another limitation is the fact that the algorithm is limited to what is written in the text.
 332 For instance, in Fig. 7, the token ‘k’, ‘Ks’ and ‘hydraulic conductivity’, all associated with hydraulic conductivity are all
 333 extracted by the NLP algorithm as they appear in this form in the abstracts. The use of synonyms can help associate tokens
 334 with similar meaning.

335 Figure 8 shows the recall and the precision of the extracted relationships according to their labeled status. For each category
 336 (negative, neutral, positive or study), the dark color represents the proportion of relationships correctly identified by the NLP
 337 algorithm. The faded color represents the relationships wrongly classified by the NLP or not found at all. Overall, most
 338 identified relationships belong to the “study” class. Note as well the larger amount of “positive” relationships compared to
 339 “negative” which may be a manifestation of some bias in reporting positive results or at least writing them as positive

340 relationships. The precision of the NLP algorithm was high for “negative” (precision = 0.88) or “study” (precision = 1.00)
341 classes. In terms of recall, the highest score is achieved for both “positive” and “study” categories.

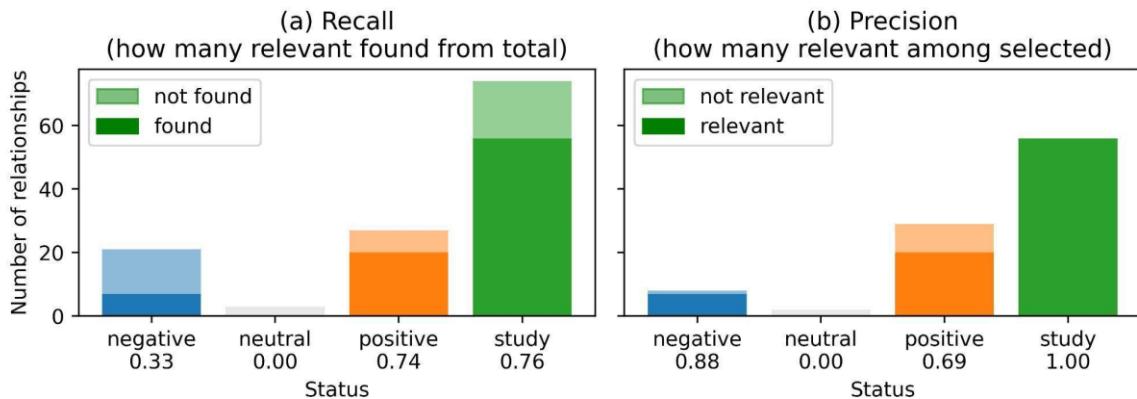


Figure 8: Recall (a) and precision (b) of classified relationships extracted from abstracts. Dark color represents the proportion of relationships correctly classified while the faded color represents relationships not found or not correctly classified. The recall and precision metric for each category is given on the X axis. Results obtained on the Meta corpus.

342
343 Based on manually labeled relationships and the ones recovered from NLP, Figure 9a offers a detailed comparison according
344 to the number of statements recovered (size of the bubble) and their correlations (colors). Such a figure has the potential to be
345 used to get a first overview of the relationships present in a large corpus of studies (e.g. for evidence synthesis). It is also
346 comparable to figures presented in the report EJP SOIL - CLIMASOMA 2022 Chapter 1. which presents a similar layout with
347 the results from the selected meta-analysis. Note that not all statements have the same relationships for specific driver/variable
348 pairs (not all studies have the same conclusions), which causes the bubbles in Fig. 9 to contain multiple colors (e.g.
349 biochar/yield, tillage/runoff). According to the relationship extraction, compost addition was positively correlated to yield,
350 residues were positively associated with lower bulk density and lower run-off, and biochar was negatively correlated to bulk
351 density and positively correlated to microbial biomass. Most of these relationships correspond well to what is reported in meta-
352 analysis (EJP SOIL - CLIMASOMA 2022, Chapter 1). As demonstrated already in Fig. 8, the NLP did not recover all
353 relationships perfectly (low recall for negative relationships) and can sometimes be completely wrong (e.g. residue/bulk
354 density). But in two thirds of all cases (66%), the relationships were classified correctly.

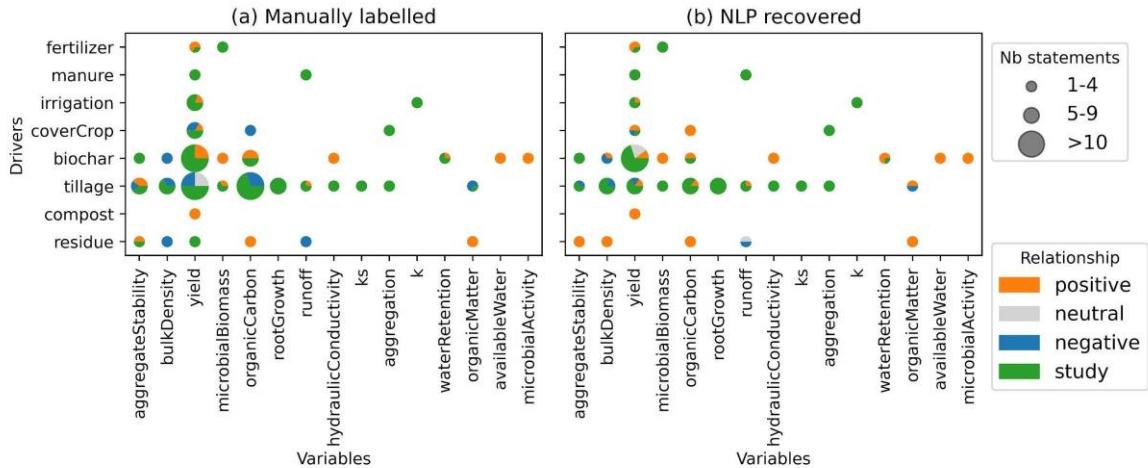


Figure 9: Relationships between drivers and variables as (a) manually labeled and (b) recovered by NLP for the Meta corpus.

355 Relationship extraction based on abstract provides a quick overview of the conclusions from a given set of documents (Fig.
 356 9). However, the classification of the extracted relationships remains a challenging task and a lot of statements just mention
 357 that the pair of drivers/variables has been studied but not the outcome of it. That is one of the limitations of the approach as
 358 not all information is contained in the abstract. Applying this technique on the conclusion part of a manuscript could help
 359 complement the relationships found.

360 In addition, to confirm that the relationships extracted are well classified, one has to manually label a given proportion of the
 361 statements found and then compare the labels with the NLP finding and iteratively improve the NLP algorithm. This procedure
 362 is tedious, but needed, as general relationships algorithms (often trained on newspaper articles or wikipedia) failed to extract
 363 meaningful relationships from field-specific scientific publications. This is in agreement with the conclusions of Furey et al.
 364 (2019). However, despite our efforts, the complexity of certain sentences (long sentences with comparison and relative clauses)
 365 was too high for our algorithm to reliably detect the relationships between a driver and a variable.

366 **4 Conclusion**

367 With the growing body of environmental scientific literature, NLP techniques can help support the needed evidence synthesis.
 368 We explored practical applications of NLP to classify documents into topics, identify knowledge gaps, build databases using
 369 regular expression and extract the main conclusion of the abstract through relationships extraction. While NLP techniques
 370 cannot replace human intervention, their automatic nature enables to quickly process a large corpus of scientific publications.
 371 When compiling an evidence synthesis, one can start by querying online search engines with specific query strings. Sets of

372 documents can then be analyzed using topic modeling and newer publications can be classified into the found topics. This
373 approach enables to identify possible knowledge gaps or topics less studied. A second step would be to extract a set of specific
374 contextual information. In this work, we demonstrated the usefulness of simple regular expressions for these tasks. Instead of
375 manually entering data into a database form, the algorithm could prefill the form for the user to verify. The database produced
376 can later be used for more quantitative analysis such as meta-analysis or machine learning techniques. Finally, a third step
377 would be to extract the main conclusion of the publications. While natural language understanding is a fastly growing field,
378 the relationship extraction algorithm developed in this work already was able to extract and classify pairs of practices (drivers)
379 and variables (soil and site properties). While their classification remains challenging and field-specific given the complexity
380 of human language, this approach already provides a good overview of the main conclusions drawn from a corpus of
381 documents.

382 Overall the NLP techniques presented in this work have practical potential to support high-throughput semi-automated
383 evidence synthesis that can be continuously updated as new publications become available. Given sufficient training data, the
384 use of more advanced methods that convert sentences to numerical vectors by the use of transformer networks (e.g. BERT,
385 Koroteev 2021), coupled with deep learning algorithms present new exciting possibilities for language understanding.

386

387 **Data availability**

388 All processing and figures presented in this manuscript are available in the form of Jupyter notebooks on the following GitHub
389 repository: <https://github.com/climasoma/nlp>. Due to copyrights restriction, the papers are not provided but a list of references
390 used is available on the GitHub repository.

391

392 **Acknowledgements**

393 This work was developed inside the CLIMASOMA project from the EJP SOIL consortium. EJP SOIL has received funding
394 from the European Union's Horizon 2020 research and innovation programme: Grant agreement No 862695 and was co-
395 funded by the involved partners: ILVO, Agroscope, SLU, WUR and CREA.

396

397 Appendices

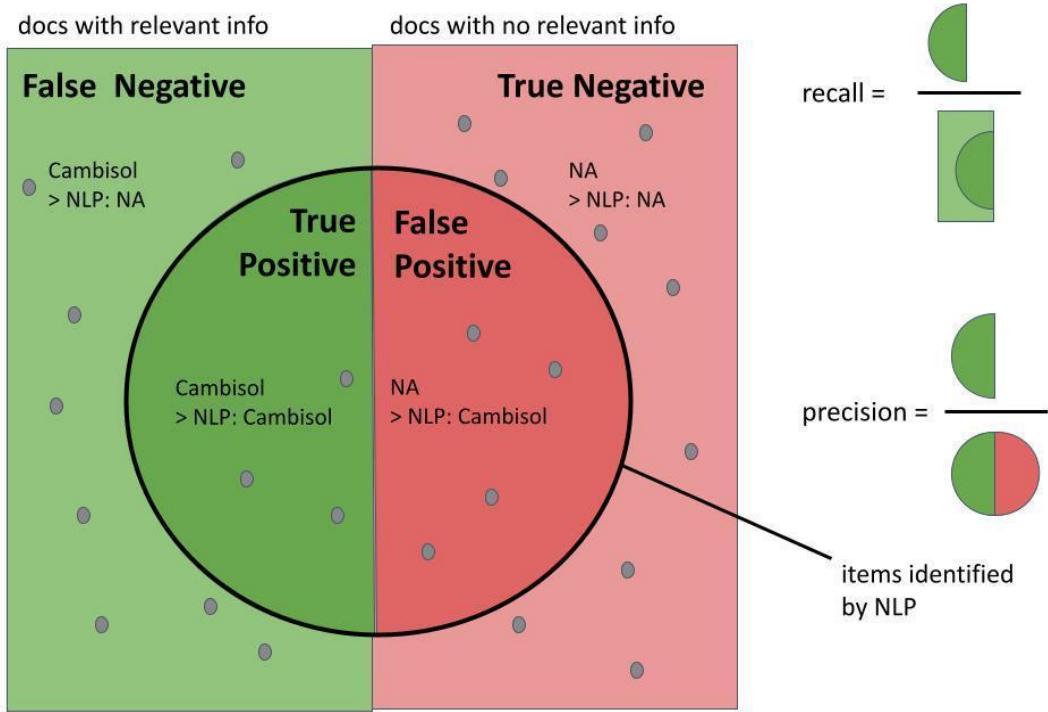


Figure A1: Schematic representation of precision and recall. Recall aims to assess how much relevant information was selected out of all the ones available in the corpus while precision aims to assess how much relevant information was in the selection.

399

400 Table A2: List of drivers and variables used in the relationships extraction.

Drivers	Variables
agroforestry	aggregate stability
biochar	aggregation
catch crop	available water
compaction	bulk density
cover crop	earthworm activity
fertilizer	earthworm biomass
intercropping	faunal activity

irrigation	faunal biomass
liming	hydraulic conductivity
compost	infiltration
manure	infiltration rate
residue	K
tillage	K(h)
traffic	K0
	Ks
	macroporosity
	microbial activity
	microbial biomass
	organic carbon
	organic matter
	penetration resistance
	rainwater penetration
	root biomass
	root depth
	root growth
	runoff
	soil strength
	water retention
	yield

401

402

403

404 **References**

- 405 Angeli, G., Johnson Premkumar, M. J., and Manning, C. D.: Leveraging Linguistic Structure For Open Domain Information
406 Extraction, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th
407 International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Proceedings of the 53rd Annual
408 Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language
409 Processing (Volume 1: Long Papers), Beijing, China, 344–354, <https://doi.org/10.3115/v1/P15-1034>, 2015.
410 Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., and Keizer, J.: The AGROVOC linked
411 dataset, 4, 341–348, 2013.

- 412 EJP SOIL - CLIMASOMA: CLIMASOMA | Final report Climate change adaptation through soil and crop management:
413 Synthesis and ways forward, 2022.
- 414 Furey, J., Davis, A., and Seiter-Moser, J.: Natural language indexing for pedoinformatics, Geoderma, 334, 49–54,
415 <https://doi.org/10.1016/j.geoderma.2018.07.050>, 2019.
- 416 Haddaway, N. R., Callaghan, M. W., Collins, A. M., Lamb, W. F., Minx, J. C., Thomas, J., and John, D.: On the use of
417 computer-assistance to facilitate systematic mapping, 16, e1129, <https://doi.org/10.1002/cl2.1129>, 2020.
- 418 Hirschberg, J. and Manning, C. D.: Advances in natural language processing, 7, 2019.
- 419 Honnibal, M. and Montani, I.: spaCy 2: Natural language understanding with bloom embeddings, convolutional neural
420 networks and incremental parsing, 2017.
- 421 Jarvis, N., Koestel, J., Messing, I., Moeys, J., and Lindahl, A.: Influence of soil, land use and climatic factors on the hydraulic
422 conductivity of soil, Hydrol. Earth Syst. Sci., 17, 5185–5195, <https://doi.org/10.5194/hess-17-5185-2013>, 2013.
- 423 Koroteev, M. V.: BERT: A Review of Applications in Natural Language Processing and Understanding, 2021.
- 424 Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J.: BioBERT: a pre-trained biomedical language
425 representation model for biomedical text mining, Bioinformatics, 36, 1234–1240,
426 <https://doi.org/10.1093/bioinformatics/btz682>, 2020.
- 427 Lin, J.: Divergence measures based on the Shannon entropy, 37, 145–151, <https://doi.org/10.1109/18.61115>, 1991.
- 428 Loper, E. and Bird, S.: NLTK: The Natural Language Toolkit, 2002.
- 429 Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D.: The Stanford CoreNLP Natural Language
430 Processing Toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System
431 Demonstrations, Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System
432 Demonstrations, Baltimore, Maryland, 55–60, <https://doi.org/10.3115/v1/P14-5010>, 2014.
- 433 Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W.: Natural language processing: an introduction, J Am Med Inform
434 Assoc, 18, 544–551, <https://doi.org/10.1136/amiajnl-2011-000464>, 2011.
- 435 Nasar, Z., Jaffry, S. W., and Malik, M. K.: Information extraction from scientific articles: a survey, Scientometrics, 117, 1931–
436 1990, <https://doi.org/10.1007/s11192-018-2921-5>, 2018.
- 437 Niklaus, C., Cetto, M., Freitas, A., and Handschuh, S.: A Survey on Open Information Extraction, 2018.
- 438 Padarian, J. and Fuentes, I.: Word embeddings for application in geosciences: development, evaluation, and examples of soil-
439 related concepts, 5, 177–187, <https://doi.org/10.5194/soil-5-177-2019>, 2019.
- 440 Padarian, J., Minasny, B., and McBratney, A. B.: Machine learning and soil sciences: a review aided by machine learning
441 tools, 6, 35–52, <https://doi.org/10.5194/soil-6-35-2020>, 2020.
- 442 Ramakrishnan, C., Patnia, A., Hovy, E., and Burns, G. A.: Layout-aware text extraction from full-text PDF of scientific articles,
443 Source Code Biol Med, 7, 7, <https://doi.org/10.1186/1751-0473-7-7>, 2012.
- 444 Rastan, R., Paik, H.-Y., and Shepherd, J.: TEXUS: A unified framework for extracting and understanding tables in PDF
445 documents, Information Processing & Management, 56, 895–918, <https://doi.org/10.1016/j.ipm.2019.01.008>, 2019.

- 446 Röder, M., Both, A., and Hinneburg, A.: Exploring the Space of Topic Coherence Measures, in: Proceedings of the Eighth
447 ACM International Conference on Web Search and Data Mining, WSDM 2015: Eighth ACM International Conference on
448 Web Search and Data Mining, Shanghai China, 399–408, <https://doi.org/10.1145/2684822.2685324>, 2015.
- 449 Sievert, C. and Shirley, K.: LDAvis: A method for visualizing and interpreting topics, in: Proceedings of the Workshop on
450 Interactive Language Learning, Visualization, and Interfaces, Proceedings of the Workshop on Interactive Language Learning,
451 Visualization, and Interfaces, Baltimore, Maryland, USA, 63–70, <https://doi.org/10.3115/v1/W14-3110>, 2014.
- 452 Tao, C., Filannino, M., and Uzuner, Ö.: Prescription Extraction Using CRFs and Word Embeddings, J Biomed Inform, 72,
453 60–66, <https://doi.org/10.1016/j.jbi.2017.07.002>, 2017.
- 454 Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., and Liu,
455 H.: Clinical information extraction applications: A literature review, Journal of Biomedical Informatics, 77, 34–49,
456 <https://doi.org/10.1016/j.jbi.2017.11.011>, 2017.