Report for BILETA March 2019.

Applicants:
1. Dr Marina Chang, (Research Centre Agroecology, Water, and Resilence, Coventry University) (member institution)
2. Dr Andelka M. Phillips, Senior Lecturer, Te Piringa Faculty of Law, University of Waikato, New Zealand - formerly Ussher Assistant Professor in Information Technology Law, Trinity College Dublin), (individual member of BILETA)
3. Dr Claudio Lombardi (Assistant Professor, School of Law, KIMEP University (formerly the Kazakhstan Institute of Management, Economics and Strategic Research), and
4. Dr I.S. Mian (Honorary Professor, Department of Computer Science, University College London).

We received a BILETA Research Award of £1000 for the project planning and development for a research project entitled **Digital Ledger Technologies and Agriculture: potential impacts and implications for whole food systems.** We are an interdisciplinary working group interested in technology assessment and regulation of current, emerging and new (bio)technologies, as well as exploring the dynamic relationships between technology, law, and agriculture. We are based in three countries – UK, Kazakhstan, and Ireland at the time of our application to you, but now New Zealand, as I (Andelka) have returned to New Zealand. Our collaborative work together is still ongoing, but we have used the funds in a slightly different way from what we had envisaged in our original proposal. We did encounter some difficulties in finding people to perform the requisite work. We are very grateful for BILETA's support, which has enabled us to develop our collaboration further.

Below is an explanation of the work we have done using the BILETA Award.

Although we initially envisaged organizing a workshop and this is something we are planning to arrange over the next year, we have used the funds to create an annotated database in GATE. This is open source and we intend to utilize the database in our future work. The task of creating the database was performed by Yucheng Ji. Below is a summary of the database work to date.

In our future work, we are exploring a number of issues. We are currently exploring other funding options to support our ongoing collaboration, so that we can proceed with the workshop in the near future. This will include discussion of issues including: the current state of affairs and future trends in "Distributed Ledger Technology" (more commonly known as blockchain technology) -- the technological infrastructure and protocols that allows simultaneous access, validation and record updating in an immutable manner across a network spread across multiple entities or locations; planned and potential applications of digital ledger technology in general as well as in agriculture; technology assessment; approaches to regulation; the possible impact of such technologies on whole food systems, small farmers, as well as societal impact more broadly; how and who will assess this technology; approaches to regulation and strategies for enforcement. I am now based in New Zealand and am planning to apply to funders in New Zealand to assist with this, but I will also be visiting the UK in May

2019, which should help us with further planning. Please also see our related papers and publications:

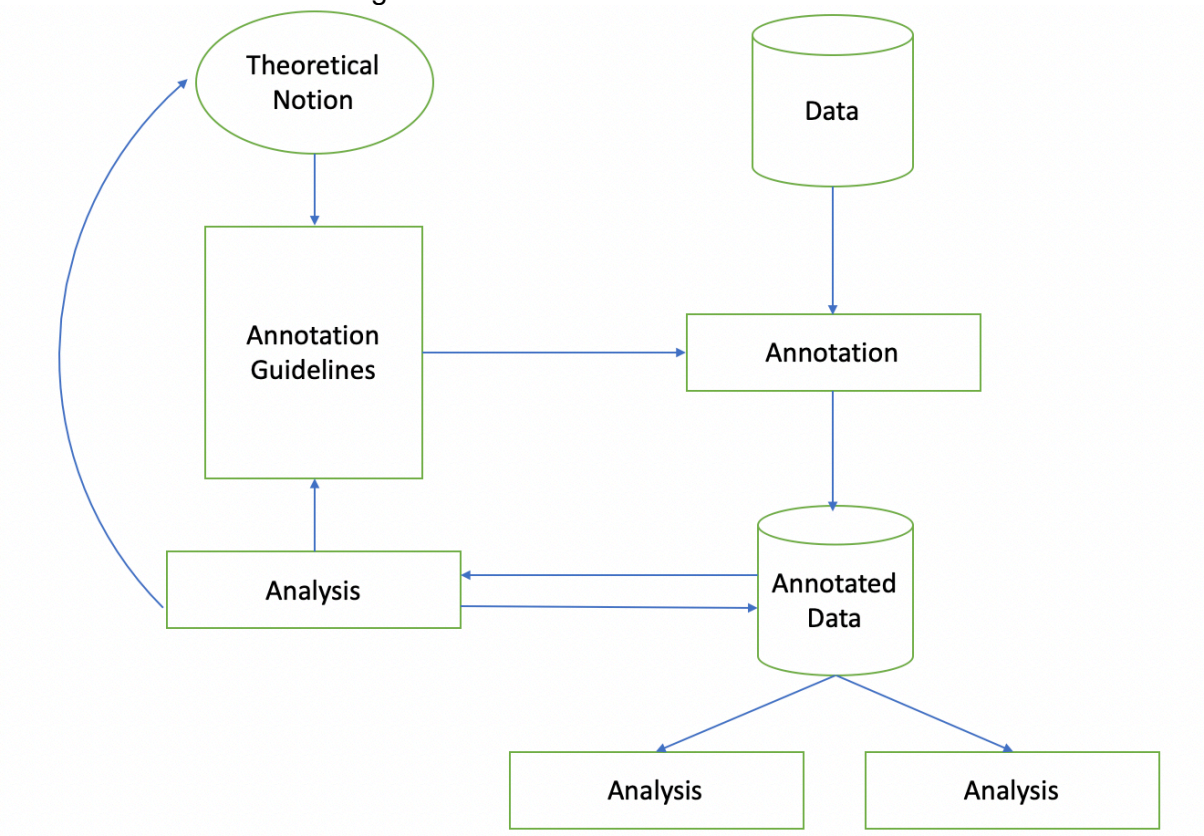
- Claudio Lombardi and Tomaso Ferrando, *EU Competition Law and Sustainability: Addressing the Broken Links*, FTAO, Brussels, 2019.
- M. Chang, C.-H. Huang, and I.S. Mian, 'A data science and historical global political ecology perspective on the financial system, agriculture and climate: from the trans-Atlantic slave trade to agroecology'. *Development*, 2018. Under review.
- Andelka M Phillips and I S Mian, 'Governance and of Future Spaces: A Discussion of Some Issues Raised by the Possibilities of Human-Machine Mergers', discussion paper for Data For Policy Government by Algorithm? Conference, London, September 6th to 7th, 2017 https://zenodo.org/record/896110#.WcDnv8iGPIU

Report for BILETA Database in GATE
Yucheng Ji

1 Introduction and Summary

This project aimed to create an annotated database of text based documents, which relate to our broader research project. Each document was annotated and the annotations included: title; abstract; date etc. Further details are listed in the description of the database schema. The overall process of creating the database could be referred to the flow chart in Figure 1

Figure 1: Flow chart for the annotations



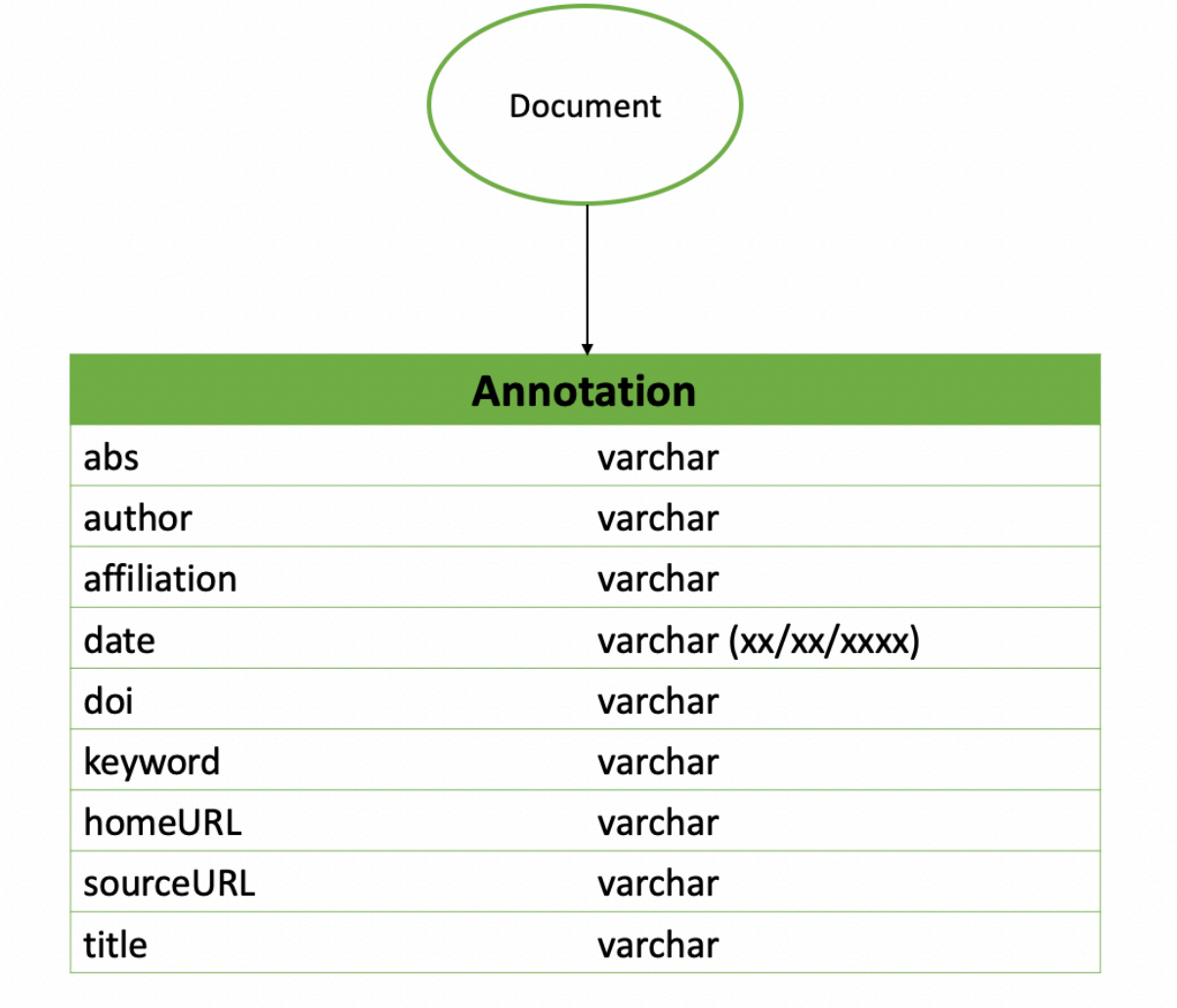
In order to allow for saving documents with different formats, such as web pages, journal papers, and pdf as well as to allow for storing meta data (annotations) for each document, we decided to store the collection of documents in an open source application, GATE [GATE]. GATE excels at text processing and analysis, which could be used to process text from the links included and also allow for saving all of the documents in our collection. Furthermore, the data collection in GATE can also be exported in different format and input to other text annotation tools (e.g. brat[brat]) and machine learning tools (e.g. weka[weka] and mallet[mallet]).

2 How to Create database in GATE

At the outset, we had hundreds of links to documents, which were often in different formats. The three main types of documents were: journal papers; *pdf* files; online articles; or web pages. Through creating a new GATE document for each text document in our collection of

links, GATE would then load the text into the application. The annotations for the documents could then be added as resource features at the bottom left of the application window. Eventually a new document is created after we save the GATE document to a specific data store, for instance *BILETA_GATE* in our case. In Figure 2, we summarize the schema of the database in GATE for all of the annotations we have, the details of how to populate each entry will be explained in the next section.

Figure 2: Schema for the GATE database



3 How to populate annotations for each document in GATE

Following the GATE database Schema in Figure 2, the population method for each annotation is explained as below:

- **abs**: abstract of the document (nan if not exist)
- **author**: author of the document (if multiple authors exist, it will be saved as author1, author2, etc.)
- **affiliation** affiliation of the corresponding author (if multiple affiliation exists for the same author, it will be saved as affiliation1_1, affiliation1_2, etc.)

Note: the affiliation follows the same structure (seperate by ';'): [institution; address; city; postcode; country]. Specifically in the institution part, if different level exist, they will be seperate by '—',e.g. [school— department— university]

- **date**: date of publishing in format [xx/xx/xxxx]

e.g. 03/12/2018 for 3rd of Dec 2018

- **doi**: doi of the document (nan if not exist)
- **homeURL**: the home page link of the documents

e.g. abstract page for a paper in arxiv

- **keyword**: the keywords of the document (nan if not exist)
- **sourceURL**: the source link of the documents

e.g. the pdf file of the paper

- **title**: the title of the document

4 Summary statistics for the current data in GATE

In the current collection (as of 28/12/2018), we have 415 documents stored in GATE. There are at least 9 annotations in each document, except for the case of those with multiple authors or multiple affiliations for the same author. By using the exportation function in GATE, the GATE document can be saved as *gatexml* and *inlinexml* files. To be specific, *gatexml* could store all of the annotations as features into a *xml* file and the features could be extracted if we parse the *xml* using scripting language like *python*. On the other hand, inline xml is similar to the txt file, which can be easily converted by simply renaming the file. Overall there are 415 *gatexml* files and 415 *inlinexml* files for each document in our data collection.

5 Exportation

5.1 Export data from GATE to Weka

Using the python script in ipyhton file, *data_exp_from_xml.ipynb*', we were able to extract the annotations for each document into a data frame in the *pandas* package. The data frame could be saved into *CSV* file using *pandas.DataFrame.to_csv* function. To create the input file to be used in Weka, we could adjust the *CSV* file to add attributes, such as name and data type to each annotation and save it to a *arff* file. For example, in Figure 3 we created an example *arff* file with three documents. Finally, the *arff* file could be loaded into machine learning tool Weka [weka] for data mining.

Figure 3: Exmaple *arff* file for Weka data mining

```
@relation annotation

@attribute abs string
@attribute author string
@attribute affiliation string
@attribute date string
@attribute doi string
@attribute homeURL string
@attribute keyword string
@attribute souceURL string
@attribute title string


% abs,author,affiliation,date,doi,homeURL,keyword,sourceURL,title

@data
,JACK BWANA,nan,22/04/2018,,,nan ,https://www.businessdailyafrica.com/analysis/columnists/How-blockchain-technology-
can-save-our-forests/4259356-4493520-rh4jaqz/,How blockchain technology can save our forests
,Suzanne Barlyn,nan,17/07/2017,,,nan,https://www.reuters.com/article/us-cyber-lloyds-report-idUSKBN1A20AB,Global cyber
attack could spur $53 billion in losses: Lloyd's of London
"Bitcoin, the digital cryptocurrency, has been celebrated as the future of money on the Internet. Although Bitcoin does
present several forward-looking innovations, it also integrates a very old concept into its digital architecture: the
mining of precious metals. Even though Bitcoin explicitly invokes mining as a metaphor and gold as an example for
understanding the cryptocurrency, there has been little critical work on the connections between Bitcoin and previous
metalist currency regimes. The following essay proposes a historical comparison with colonial South American silver
mining and the global currency regime based on the New World silver peso it created as a way to interrogate Bitcoin.
The comparison with colonial South America, and specifically the silver mining economy around the Cerro Rico de Potosí,
will help to develop a historical and political understanding of Bitcoin's stakes, including questions of resources,
labor, energy, and ecology. Mining and the extractive apparatus that accompanies it always imply massive-scale
earthworks that reshape the planet itself, a process known as terraforming. The Potosí comparison will reveal Bitcoin
to form part of a similar process of digital primitive accumulation we can provisionally name cryptoforming.",Zac
Zimmer,UC Santa Cruz; nan; nan; nan; nan,xx/04/2017,10.1353/tech.2017.0038,https://muse.jhu.edu/article/
662979,nan,https://muse.jhu.edu/article/662979/pdf?casa_token=-4hU31Ff8-EAAAAA:
295xvuNUSIciPoB2ojuh5zNLPrfzTIDxaivygdJMECtCAvy8js7-A3fk7l79zleoAaUXdvHgTA,Bitcoin and Potosí Silver: Historical
Perspectives on Cryptocurrency
```

5.2 Export data from GATE to Mallet

For the mallet tool, we could save the collection of data into a collected *.mallet* file

For command tool in Windows version:

bib\mallet import-dir --input data_directory --output collection_name.mallet
--keep-sequence --remove-stopwords

For terminal in Mac version:

./bin/mallet import-dir --input data_directory --output collection_name.mallet
--keep-sequence --remove-stopwords

(Where we could replace 'data_directory' as the path to the folder containing the collection of documents (the *txt* files) and 'collection_name' as the name for the collection of text documents (for example, *BILETA*) in both windows and mac version.)

5.3 Export data from GATE to brat

For the brat text analysis tool, the input data could be created based on the *inlinexml* file we exported from the GATE database. Brat need two files to load one document, which are the text documents stored in *txt* files and annotation documents store in the *.ann* files. For the *txt* file, we could convert the *inlinexml* directly into *txt* file by renaming it. Where no annotation existed, we were able to create an empty *.ann* file to be loaded into brat.

References

[] Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V., 2002, July. GATE: an architecture for development of robust HLT applications. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 168-175). Association for Computational Linguistics.

[] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. and Tsujii, J.I., 2012, April. BRAT: a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 102-107). Association for Computational Linguistics.

[] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), pp.10-18.

[] McCallum, A.K., 2002. Mallet: A machine learning for language toolkit.