Blockchain Lexicography: Prototyping the Collaborative, Participatory Post-dictionary

Daniel McDonald, Eveline Wandl-Vogt
Bitpanda GmbH; Austrian Academy of Sciences, Austrian Centre for Digital Humanities,
Austrian Academy of Sciences, Austrian Centre for Digital Humanities
Burggasse 116/2 1070 Vienna
Wohllebengasse 12-14/2 1040 Vienna
daniel.mcdonald@bitpanda.com
tEveline.Wandl-Vogt@oeaw.ac.at

Abstract

In this paper, the authors introduce blockchain lexicography, developed and prototyped within the framework of the open innovation exploration space (Research Group Methods and Innovation) at the Austrian Academy of Sciences. Blockchain lexicography exploits emerging technologies (e.g. the blockchain), social developments (do-it-yourself-science, crowd-innovation) and management methods and practices (open innovation), applying them to a case study of lexicography in order to create an accessible, constantly evolving linguistic resource. The authors deliver the design and a prototype of the system, as well as related data. The system, *wugsy*, asks users to provide natural-language texts for images, to score others' texts, or to assess the accuracy of tag clouds, with text types (i.e. stories, descriptions, etc.) tailored to match user profiles. Answers are recorded in a distributed database, which can be hosted, verified or queried by anyone. Responses to games are scored by consensus and rewarded proportionally with a cryptocurrency token. A simple API allows extraction and filtering of database contents.

Keywords: blockchain, consensus, trust, democratisation, post-dictionary

1. Context

The increasing pervasiveness of digital communication in everyday life facilitates the development and use of novel computational methods that allow better understanding language, society and culture. Lexicography is one area of research that stands to benefit from increasingly digitised life, in terms of *(a)* research presentation; *(b)* use of social media and digital news as corpora, *(c)* interlinking and harmonisation of linguistic data; and *(d)* opening up communicative channels between experts and volunteers (Chesbrough 2006). Digital lexicography, however, has so far rarely made use of the affordances of new media, making it difficult to imagine the future of lexicography; as Hanks explains, it is currently still 'too early, to say, which form innovative dictionaries of the future will take' (2012, p. 82). For this reason, exploration of emerging technologies for the purposes of uncovering new ways of building lexicographical resources is timely.

A parallel computational development is the blockchain (Wood 2014, Pilkington 2015)—a decentralised, trustless ledger that can accurately keep track of digital information. To date, the most common use-case for blockchain technology is as a currency or payment network (e.g. Bitcoin, Ethereum). Recently, however, a number of blockchain research projects have aimed to go beyond cryptocurrency applications, using blockchains as ways of providing proof of existence of documents, as well as tracking migration and medical histories. Blockchain-based systems permit the transfer of real or symbolic value in a way that is very resilient to system outages and malicious code. Meanwhile, blockchain-based databases are provably open-source, limiting researcher bias, increasing reproducibility, and promoting data re-use. For this reason, blockchains have a key potential use case within the open source, open science and open innovation movements, which aim to facilitate access to research tools, data and publications. While cryptocurrency systems have demonstrated the utility of blockchains as both a reward mechanism and store of value, still to be empirically tested is the suitability of blockchain protocols for research data collection.

Related to both the increased presence of digital communication and the rise of decentralised networks is crowdsourcing—the targeted collection of large amounts of data from a pool of online participants. Though some crowdsourcing work in linguistics has been criticised based on the accuracy of generated results, as well as issues of exploitation of labour, ethical crowdsourcing is a major component within the emerging framework of open innovation (Sloane 2007, Chesbrough 2006), due to the fact that crowdsourcing engages the public in science and research, promoting democratisation and the synergy of diverse sources and kinds of knowledge.

Blockchains provide a natural, but thus far underutilised, complement to crowdsourcing tasks. By storing data and rewards in a publicly accessible database that is very difficult to corrupt, it is possible to develop crowdsourcing systems that are provably fair, with results that are inherently publicly accessible. We therefore believe that the combination of blockchain technology and crowdsourcing methods can lead to systems for natural language data generation and collection that surpass current methods in terms of both utility and fairness.

2. Aim

In this paper, the authors describe the potential for emerging technologies to be put to use in the context of the post-dictionary phenomenon at the currently founded *exploration space @ ÖAW* (the Austrian Academy of Sci-

ences). They introduce the concept of blockchain lexicography and offer *wugsy*, an initial, open-source prototype of such a system, with the aim of furthering knowledge discovery in the context of linguistic, biological and cultural diversity. The open-source platform gives linguistic tasks to a crowd, and stores the results of these tasks within a blockchain. A separate, but related chain, distributes rewards to participants based on emerging consensus regarding the quality of their answers. Because the data accumulated by the system is free to access, its downstream applications are many. For our purposes, however, we aim to demonstrate that the system can generate insights that are novel and appropriate for inclusion within a dynamically generated post-dictionary.

3. Prototype

wugsy is human-centred, devised against a background of design thinking (Plattner, Meinel and Weinberg 2009) and agile development. Via a web platform (implemented in Python 3/Django), images are presented to actors, alongside one of a number of possible tasks. The user may variously be asked to:

1. Write a natural language text related to the image
2. Score/rank another user's existing text
3. Select relevant terms that appear within a visualised tag cloud generated through a simple NLP pipeline run over a text
4. Score/rank the accuracy of a tag cloud

The languages and text types requested from users can vary based on current gaps in the dataset and on users' stated language proficiencies, interests and areas of expertise. Tag clouds are generated by parsing texts with `spaCy`, and using POS tags and dependency positions and NER to identify likely tags. Results from these different games (i.e. natural language content, selected tags, rankings of others' stories and tag selections) are then sent to a decentralised database (McConaghy et al. 2016) hosted by those who wish to use the data for downstream tasks. As other actors score the accuracy of stories and selected tags, it becomes possible to determine answer quality by consensus. The degree of consensus for a given question dictates the size of the reward for an individual answer. Actor history can be used to further scale the size of the reward, and incentivise high-quality or high-effort answers (e.g. short composition or brainstorming tasks). Rewards are released to each user's account in the form of an Ethereum-based ERC20 token, which could be given an intrinsic, fluctuating value derived from, e.g., real-world investment in the infrastructure, through fees for API calls to nodes that host the database, or through fees paid in order to add new kinds of data and questions to the crowd. Such a structure incentivises not only participation in games, but also the addition of new data, which expands the explanatory potential of the project, and the hosting of nodes, which play an important role in the overall security and stability of the network.

4. Workflow

Taking lexicography as an aim, the workflow for the system is fairly simple. Europeana's historical multimedia collection (Haslhofer and Isaac 2011) is used as an initial image and caption dataset, with users asked to variously generate texts about images, score others' texts, or score the accuracy of tag clouds. These small, compartmentalised tasks are provided by a dynamic visualisations within a web front-end; the combined use of scores, currency rewards and high-quality visualisation of natural language text each gamify the process of data collection, motivating users to produce high-quality content. Participation in games can be anonymous, but participants are rewarded for adding user profiles, because the coupling of profile and answer data makes possible both targeted questioning, and, downstream, more nuanced insights into language use in different dialects, registers and demographics.

An open-source API allows querying the generated data, and dynamically presenting interesting insights online in real-time. The potential use of the API for lexicographic tasks is explored: searching information from the generated tag clouds gives us an insight into relationships between particular words, images and narratives; by restricting search results based on users' overall scores, we can see the differences between high and low-quality submissions, and consider their implications for the design of novel kinds of dictionary. Similarly, we explore how queries containing location filters can be used to uncover regional variations.

5. Design Parameters

The codebase is designed with five key design parameters in mind. Namely, the developed system is:

(a) inherently multilingual

(b) responsive to user-specific expertise

(c) self-improving

(d) adaptable to new kinds of language tasks

(e) sensitive to practices of open innovation and open science

Regarding parameters *(a)* and *(b)*, rewards are scaled by the current size of a given language's dataset, with profiles of crowdsourcing participants used to present language problems to participants in line with their stated interests and areas of expertise. Such a design means that languages and content areas with less accumulated information can be prioritised by a relative increase in reward sizes, and by putting more questions from less popular languages and content areas to the user base.

Regarding parameter *(c)*, the authors aim to use the incoming streams of crowdsourced answers continually to train algorithms responsible for selecting problems that are served to the crowd. For example, the algorithms that transform users' texts into tag clouds can be refined based on the kinds of tags that users mark as accurate, or by users' scoring of the tag clouds themselves.

Regarding parameter *(d)*, within the early prototype, lexicography acts as a test-case for a more abstract system that

is equally well-suited to other areas of research. By using different kinds of initial datasets, and by developing new kinds of language games, we expect the system to be able to collect data suitable for use in diverse kinds of research, including linguistic typology (in classifying languages and dialects), computational linguistics (i.e. in natural language generation and parsing), and the social, political and population sciences (in mapping language use to demographic details, or uncovering attitudes toward the data shown to participants).

Regarding parameter *(e)*, the prototype described here not only facilitates novel kinds of research, but, in doing so, also necessarily commits to core values of open science and innovation. *wugsy* guarantees open data and open-source development, connects problems with those best capable of solving them, and thus promotes the creation of knowledge that is provably accessible and diverse. Furthermore, *wugsy* empowers marginalised actors: because the proposed system is multilingual, and because rewards are scaled to incentivise answers for domains in which less data has accumulated, global participants can potentially receive fair compensation for their work.

6. Bibliographical References

Chesbrough, Henry W. (2006): Open Innovation. The New Imperative for Creating and Profiting from Technology. Boston.

Fellbaum, Christiane (2014): Large-scale Lexicography in the Digital Age. International Journal of Lexicography, Volume 27, Issue 4, 1 December 2014, Pages 378–395, https://doi.org/10.1093/ijl/ecu018 (accessed: 07.01.2018).

Hanks, Patrick (2012): Lexicography from earliest times to the present. In: Allan, K. (Ed.): The Oxford Handbook of the History of Linguistics. http://www.patrickhanks.com/uploads/5/1/4/9/5149363/2012dlexicography_from_earliest_times.pdf (accessed: 07.01.2018).

Haslhofer, B., and Isaac, A. (2011). *data.europeana.eu*: The europeana linked open data pilot. In International Conference on Dublin Core and Metadata Applications (pp. 94-104).

McConaghy, T., Marques, R., Müller, A., De Jonghe, D., McConaghy, T., McMullen, G., Henderson, R., Bellemare, S. and Granzotto, A., (2016). BigchainDB: a scalable blockchain database. white paper, BigChainDB. https://docs.bigchaindb.com/en/latest/

Pilkington, Marc (2015). Blockchain Technology: Principles and Applications. Research Handbook on Digital Transformations, edited by F. Xavier Olleros and Majlinda Zhegu. Edward Elgar, 2016. https://ssrn.com/abstract=2662660 (accessed: 07.01.2018).

Plattner, Hasso, Meinel, Christoph, and Weinberg Ulrich (2009): Design Thinking. München.

Sloane, Paul (Ed.; 2011): A Guide to Open Innovation and Crowdsourcing: practical tips, advice and examples from leading experts in the field. https://books.google.at/books?hl=de&lr=&id=mscjeFHY8NQC&oi=fnd&pg=PR4&dq=open+innovation+crowdsourcing&ots=2Nku9jDmRC&sig=L9CiT3ILk2plCtVmWn4YmGKca7w#v=onepage&q=open%20innovation%20crowdsourcing (accessed: 07.01.2018).

Wood, G., (2014). Ethereum: A secure decentralised generalised transaction ledger. Ethereum Project Yellow Paper, 151.