Digging in documents: using text mining to access the hidden knowledge in Dutch archaeological excavation reports

Brandsen, A.

Bibliography

Abrahamse, J., Blom, A., Bouwmeester, H., Bos, J., Brinkkemper, O., Brounen, F., Cohen, K., Dambrink, R., Bruijn, R.d., Groot, T.d., Kort, J.d., Vries, F.d., Vries, S.d., Eerden, M., Erkens, G., Feiken, H., Gouw-Bouman, M., Groenewoudt, B., Hijma, M., Huisman, D., Jansen, B., Kosian, M., Koster, K., Kriek, M., Lascaris, M., Lauwerier, R., Maas, G., Marges, V., Pierik, H., Rensink, E., Romeijn, E., Schokker, J., Smit, B., Snoek, M., Speleers, B., Stafleu, J., Theunissen, E., Beek, R.v., Jagt, I.v.d., Doesburg, J.v., Reuler, H.v., Vos, P. & Weerts, H. (2017). Knowledge for Informed Choices. Tools for more effective and efficient selection of valuable archaeology in the Netherlands. In R. Lauwerier, M. Eerden, B. Groenewoudt, M. Lascaris, E. Rensink, B. Smit, B. Speleers & J.v. Doesburg, eds., *Nederlandse Archeologische Rapporten*, volume 55. Rijksdienst voor het Cultureel Erfgoed, Amersfoort. ISBN 9789057992773.

Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*. Pan Books (UK). ISBN 0-330-25864-8.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S. & Vollgraf, R. (2019). FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59. Association for Computational Linguistics, Minneapolis, Minnesota. DOI: 10.18653/v1/N19-4010.

Akhtyamova, L. (2020). Named Entity Recognition in Spanish Biomedical Literature: Short Review and Bert Model. In *26th Conference of Open Innovations Association (FRUCT)*, pp. 1–7. IEEE Computer Society, Yaroslavl, Russia. ISBN 9789526924427. ISSN 23057254. DOI: 10.23919/FRUCT48808.2020.9087359.

Amrani, A., Abajian, V. & Kodratoff, Y. (2008). A chain of text-mining to extract information in archaeology. In *Information and Communication Technologies: From Theory to Applications, ICTTA 2008.*, pp. 1–5. Damascus, Syria. DOI: 10.1109/ICTTA.2008.4529905.

Annaert, R. (2018). *Het Vroegmiddeleeuwse Grafveld Van Broechem, Volume II Analyse*. Habelt Verlag, Bonn.

Athens, J.S. (1993). Cultural resource management and academic responsibility in archaeology: A further comment. *SAA Bulletin*, 11(2), pp. 6–7.

Auger, C.P.C.P. (1975). *Use of reports literature*. Archon Books. ISBN 020801506X.

Auger, C.P.C.P. (1989). *Information sources in Grey literature*. Bowker-Saur. ISBN 0862918715.

Averett, E.W., Counts, D. & Gordon, J. (2016). *Mobilizing the past for a digital future: the potential of digital archaeology*. Technical report, Creighton University. DOI: 10.17613/M6HJ56.

Barbour, R. (2018). *Doing focus groups*. Sage, Los Angeles London. ISBN 9781473912441.

Bartalesi, V., Meghini, C., Metilli, D. & Andriani, P. (2016). Usability Evaluation of the Digital Library DanteSources. In *International Conference on Theory and Practice of Digital Libraries*, pp. 191–203. Springer, Cham. DOI: 10.1007/978-3-319-39513-5_18.

Bazelmans, J., Brinkkemper, O., Deeben, J., Van Doesburg, J., Lauwerier, R. & Zoetbrood, P. (2005). Mag het ietsje meer zijn? Een onderzoek naar de door bedrijven opgestelde Programma's van Eisen voor archeologisch onderzoek uit de periode 2003-2004. *Rapportage Archeologische Monumentenzorg*, 120.

Beck, K., Beedle, M., Bennekum, A.V., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B.,

Martin, R.C., Mellor, S., Schwaber, K., Sutherland, J. & Thomas, D. (2001). Manifesto for Agile Software Development. http://agilemanifesto.org/.

Behnert, C. & Lewandowski, D. (2017). A framework for designing retrieval effectiveness studies of library information systems using human relevance assessments. *Journal of Documentation*, 73(3), pp. 509–527. ISSN 00220418. DOI: 10.1108/JD-08-2016-0099.

Beltagy, I., Lo, K. & Cohan, A. (2020). SCIBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, Hong Kong, China. ISBN 9781950737901. DOI: 10.18653/v1/d19-1371.

Bennett, R., Cowley, D. & De Laet, V. (2014). The data explosion: Tackling the taboo of automatic feature recognition in airborne survey data. *Antiquity*, 88(341), pp. 896–905. ISSN 0003598X. DOI: 10.1017/S0003598X00050766.

Bevan, A. (2015). The data deluge. *Antiquity*, 89(348), pp. 1473–1484. DOI: 10.15184/aqy.2015.102.

Bloomberg, J. (2013). The Big Data Long Tail. http://www.devx.com/blog/the-big-data-long-tail.html.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5(1), pp. 135–146.

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Graeme, L., O'Neill, O., Rawlins, M., Thornton, J., Vallance, P. & Walport, M. (2012). *Science as an open enterprise. The Royal Society Science Policy Centre report 02/12*. Technical Report June, The Royal Society, London. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf.

Boyd, D. & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*. ISSN 1369118X. DOI: 10.1080/1369118X.2012.678878.

Bradley, R., Haselgrove, C., Vander Linden, M. & Webley, L. (2016). *The later prehistory of north-west Europe: The evidence of development-led fieldwork*. Oxford University Press, Oxford. ISBN 9780199659777.

Bramer, W.M., Giustini, D., Kramer, B.M. & Anderson, P. (2013). The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. *Systematic reviews*, 2(1), p. 115. ISSN 20464053. DOI: 10.1186/2046-4053-2-115.

Branco, P., Torgo, L. & Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions.

Brandsen, A. (2018). alexbrandsen/archaeo-CRF: Version 0.1 of Archaeo CRF. *Zenodo Repository*. DOI: 10.5281/ZENODO.1238861.

Brandsen, A. (2019). alexbrandsen/dutch-archaeo-NER-dataset: First version. *Zenodo Repository*. DOI: 10.5281/ZENODO.3544544.

Brandsen, A. (2020). alexbrandsen/archaeo-document-classification-dataset: Second version. *Zenodo Repository*. DOI: 10.5281/ZENODO.4115747.

Brandsen, A. (2021a). Archaeological entities and timespans extracted from all archaeology documents available in DANS EASY in 2017. DOI: 10.17026/dans-zcs-7b72.

Brandsen, A. (2021b). ArcheoBERTje - A Dutch BERT model for the Archaeology domain. *Zenodo Repository*. DOI: 10.5281/zenodo.4739063.

Brandsen, A. & Koole, M. (2021). Labelling the Past: Data Set Creation and Multi-label Classification of Dutch Archaeological Excavation Reports. *Language Resources and Evaluation*. DOI: 10.1007/s10579-021-09552-6.

Brandsen, A., Lambers, K., Verberne, S. & Wansleeben, M. (2019). User Requirement Solicitation for an Information Retrieval System Applied to Dutch Grey Literature in the Archaeology Domain. *Journal of Computer Applications in Archaeology*, 2(1), pp. 21–30. DOI: 10.5334/jcaa.33.

Brandsen, A. & Lippok, F. (2021). AGNES case study data. *Zenodo Repository*. DOI: 10.5281/zenodo.4737564.

Brandsen, A., Verberne, S., Lambers, K. & Wansleeben, M. (2021a). Can BERT Dig It? - Named Entity Recognition for Information Retrieval in the Archaeology Domain. *arXiv*. http://arxiv.org/abs/2106.07742.

Brandsen, A., Verberne, S., Lambers, K. & Wansleeben, M. (2021b). Usability Evaluation for Online Professional Search in the Dutch Archaeology Domain. *arXiv*. http://arxiv.org/abs/2103.04437.

Brandsen, A., Verberne, S., Wansleeben, M. & Lambers, K. (2020). Creating a Dataset for Named Entity Recognition in the Archaeology Domain. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 4573–4577. European Language Resources Association, Marseille, France. https://www.aclweb.org/anthology/2020.lrec-1.562/.

Brandt, R., Drenth, E., Montforts, M., Proos, R., Roorda, I. & Wiemer, R. (1992). *Archeologisch Basisregister*. Technical report, Rijksdienst voor Cultureel Erfgoed, Amersfoort.

Bulatovic, N., Gnadt, T., Romanello, M., Stiller, J. & Thoden, K. (2016). Usability in digital humanities - Evaluating user interfaces, infrastructural components and the use of mobile devices during research process. In N. Fuhr, L. Kovács, T. Risse & W. Nejdl, eds., *Research and Advanced Technology for Digital Libraries. TPDL 2016. Lecture Notes in Computer Science*, volume 9819 LNCS, pp. 335–346. Springer, Cham. ISBN 9783319439969. ISSN 16113349. DOI: 10.1007/978-3-319-43997-6_26.

Byrne, K. & Klein, E. (2010). Automatic Extraction of Archaeological Events from Text. In B. Frischer, J. Crawford, & D. Koller, eds., *Making History Interactive: Computer Applications and Quantitative Methods in Archaeology 2009*, pp. 48–56. BAR International Series 2079, Oxford.

Capannini, G., Nardini, F.M., Perego, R. & Silvestri, F. (2011). Efficient diversification of web search results. In *Proceedings of the VLDB Endowment*, pp. 451 – 459. Seattle, Washington. ISSN 21508097. DOI: 10.14778/1988776.1988781.

Carpineto, C. & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1). ISSN 03600300. DOI: 10.1145/2071389.2071390.

Chan, B., Schweter, S. & Möller, T. (2021). German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6788–6796. International Committee on Computational Linguistics, Barcelona, Spain. DOI: 10.18653/v1/2020.coling-main.598.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. SAGE Publications Ltd, London. ISBN 0761973532. DOI: 10.1080/17482620600881144.

Cheng, X., Bowden, M., Bhange, B.R., Goyal, P., Packer, T. & Javed, F. (2020). An End-to-End Solution for Named Entity Recognition in eCommerce Search. *arXiv*. http://arxiv.org/abs/2012.07553.

Cherman, E.A., Monard, M.C. & Metz, J. (2011). Multi-label problem transformation methods: a case study. *CLEI Electronic Journal*, 14(1), p. 4.

Chowdhury, G.G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), pp. 51–89. ISSN 00664200. DOI: 10.1002/aris.1440370103.

Claeys, J., Baetsen, S., Drenth, E., Huizer, J., Jaspers, N., Kempkens, J., Lupak, T. & Melkert, M. (2012). Onder de Gelderakkers een uitgestrekte meerperiodensite in Hilvarenbeek. Een inventariserend veldonderzoek door middel van proefsleuven (IVO-P). *ADC Archeoprojecten Rapport*, 2613. DOI: 10.17026/dans-zvw-3mpv.

Cohen, L., Manion, L., Morrison, K., Manion, L. & Morrison, K. (2002). *Research Methods in Education*. Routledge. ISBN 9780203224342. DOI: 10.4324/9780203224342.

Concrete5 (2018). Concrete5 Content Management System. https://www.concrete5.org/.

Copara, J., Naderi, N., Knafou, J., Ruch, P. & Teodoro, D. (2020). Named entity recognition in chemical patents using ensemble of contextual language models. *arXiv*. ISSN 23318422. http://arxiv.org/abs/2007.12569.

Copeland, J.M. (1983). Information retrieval systems for archaeological data. In J. Haigh, ed., *Proceedings of the Conference on Computer Applications and Quantitative Methods in Archaeology (CAA) 1983*, pp. 39–45. School of Archaeological Sciences, University of Bradford, Bradford.

Corstius, H.B. (1981). *Opperlandse taal-& letterkunde*. Querido. ISBN 9789021451343.

Costopoulos, A. (2016). Digital Archeology Is Here (and Has Been for a While). *Frontiers in Digital Humanities*, 3. ISSN 2297-2668. DOI: 10.3389/fdigh.2016.00004.

Council of Europe (1992). European Convention on the Protection of the Archaeological Heritage (Revised). http://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/143.

Cowan, B., Zethelius, S., Luk, B., Baras, T., Ukarde, P. & Zhang, D. (2015). Named entity recognition in travel-related search queries. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 3935–3941. AAAI Press, Austin, Texas. ISBN 9781577357032.

Cowley, D.C. (2012). In with the new, out with the old? Auto-extraction for remote sensing archaeology. In C.R. Bostater, S.P. Mertikas, X. Neyt, C. Nichol, D. Cowley & J.P. Bruyant, eds., *Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2012*. SPIE, Edinburgh, UK. ISBN 9780819492722. ISSN 0277786X. DOI: 10.1117/12.981758.

Croft, W.B., Metzler, D. & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Pearson. ISBN 978-0136072249.

Cunningham, H., Gaizauskas, R.J. & Wilks, Y. (1995). *A general architecture for text engineering (GATE): A new approach to language engineering R & D*. University of Sheffield, Department of Computer Science.

DANS (2019). DANS EASY. https://dans.knaw.nl/en/about/services/easy.

De Mauro, A., Greco, M. & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. In *AIP Conference Proceedings*, volume 1644, pp. 97–104. American Institute of Physics Inc. ISBN 9780735412835. ISSN 15517616. DOI: 10.1063/1.4907823.

De Mauro, A., Greco, M. & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65(3), pp. 122–135. ISSN 00242535. DOI: 10.1108/LR-06-2015-0061.

De Romas, R. (2019). *Multi-label Text Classification for Ground Lease Documents*. Thesis, University of Amsterdam.

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G. & Nissim, M. (2019). BERTje: A Dutch BERT Model. *arXiv*. http://arxiv.org/abs/1912.09582.

Dejong, M. & Schellens, P.J. (1997). Reader-Focused Text Evaluation: An Overview of Goals and Methods. *Journal of Business and Technical Communication*, 11(4), pp. 402–432. DOI: 10.1177/1050651997011004003.

Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., Kouril, M., Marsolo, K. & Solti, I. (2012). Building gold standard corpora for

medical natural language processing tasks. *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012, pp. 144–153.

Delobelle, P., Winters, T. & Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3255–3265. Association for Computational Linguistics, Online. ISSN 23318422. DOI: 10.18653/v1/2020.findings-emnlp.292.

Deshmukh, A.A. & Sethi, U. (2020). IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles. *arXiv*. https://arxiv.org/abs/2007.12603.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long an, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. DOI: 10.18653/v1/N19-1423.

Dudek, D., Mastora, A. & Landoni, M. (2007). Is Google the answer? A study into usability of search engines. *Library Review*, 56(3), pp. 224–233. DOI: 10.1108/00242530710736000.

Eerden, M., Groenewoudt, B., de Groot, T., Theunissen, E. & Feiken, H. (2017). Synthesising data from development- led archaeological research. In R. Lauwerier, M. Eerden, B. Groenewoudt, M. Lascaris, E. Rensink, B. Smit, B. Speleers & J. van Doesburg, eds., *Knowledge for Informed Choices Tools for more efective and efcient selection of valuable archaeology in the Netherlands*. Rijksdienst voor het Cultureel Erfgoed, Amersfoort.

Effros, B. (2003). *Merovingian Mortuary Archaeology and the Making of the Early middle ages*. University of California Press, Berkeley.

ElasticSearch (2018). Theory Behind Relevance Scoring. https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html.

Eramian, M., Walia, E., Power, C., Cairns, P. & Lewis, A. (2017). Image-based search and retrieval for biface artefacts using features capturing archaeologically significant characteristics. *Machine Vision and Applications*, 28(1-2), pp. 201–218. ISSN 14321769. DOI: 10.1007/s00138-016-0819-x.

Esmailpour, R., Ebrahimy, S., Fakhrahmad, S.M., Mohammadi, M. & Abbaspour, J. (2019). Developing an effective scheme for translation and expansion of Persian user queries. *Digital Scholarship in the Humanities*. ISSN 2055-7671. DOI: 10.1093/llc/fqz041.

Evans, C., Tabor, J. & Vander Linden, M. (2014). Making time work: Sampling floodplain artefact frequencies and populations. *Antiquity*, 88(339), pp. 241–258. ISSN 0003598X. DOI: 10.1017/S0003598X0005033X.

Evans, T.N.L. (2015). A reassessment of archaeological grey literature: Semantics and paradoxes. *Internet Archaeology*, 40. ISSN 13635387. DOI: 10.11141/ia.40.6.

Falkingham, G. (2005). A Whiter Shade of Grey: A new approach to archaeological grey literature using the XML version of the TEI Guidelines. *Internet Archaeology*, 17. DOI: 10.11141/ia.17.5.

Farace, D.J. & Schoptimefel, J. (2010). *Grey literature in library and information studies*. De Gruyter Saur. ISBN 9783598117930. http://hal.univ-lille3.fr/hal-01288536.

Fehr, H. (2008). Germanische Einwanderung oder kulturelle Neuorientierung? Zu den Anfängen des Reihengräberhorizontes. In S. Brather, ed., *Zwischen Spätantike und Frühmittelalter*, pp. 67–102. Walter de Gruyter, Berlin. DOI: 10.1515/9783110210729.2.67.

Feldman, R. & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Montreal, Canada. http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf.

Feldman, R. & Sanger, J. (2007). *The text mining handbook : advanced approaches in analyzing unstructured data*. Cambridge University Press. ISBN 9780521836579.

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A. & James, S. (2020). Machine Learning for Cultural Heritage: A Survey. *Pattern Recognition Letters*. ISSN 01678655. DOI: 10.1016/j.patrec.2020.02.017.

Fischer, A., Londen, H.v., Bercken, A.B.v.d., Visser, R. & Renes, J. (2021). NAR 68 Urban farming and ruralisation in the Netherlands (1250 up to the nineteenth century), unravelling farming practice and the use of (open) space by

synthesising archaeological reports using text mining. *Nederlandse Archeologische Rapporten (NAR)*, 68.

Fokkens, H., Steffens, B. & van As, S. (2016). *Farmers, fishers, fowlers, hunters. Knowledge generated by development-led archaeology about the Late Neolithic, the Early Bronze Age and the start of the Middle Bronze Age (2850 - 1500 cal BC) in the Netherlands*. Rijksdienst voor het Cultureel Erfgoed, Amersfoort.

Foley, R. (1981). Off-site archaeology: an alternative approach for the short-sited. In I. Hodder, G. Isaac & N. Hammond, eds., *Pattern of the past: studies in honour of David Clarke*. Cambridge University Press, Cambridge. ISBN 9780521108430.

Gartner Glossary (2021). Definition of Big Data - Gartner Information Technology Glossary. https://www.gartner.com/en/information-technology/glossary/big-data.

Gattiglia, G. (2015). Think big about data: Archaeology and the Big Data challenge. *Archäologische Informationen*, 38(1), pp. 113–124. ISSN 2197-7429. DOI: 10.11588/ai.2015.1.26155.

Gerjets, P., Kammerer, Y. & Werner, B. (2011). Measuring spontaneous and instructed evaluation processes during Web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. *Learning and Instruction*, 21(2), pp. 220–231. DOI: 10.1016/j.learninstruc.2010.02.005.

Gibbs, F. & Owens, T. (2012). Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs. *Digital Humanities Quarterly*, 6(2). ISSN 1938-4122. http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html.

Gibbs, M. & Colley, S. (2012). Digital Preservation,: Online access and historical archaeology 'grey literature' from New South Wales, Australia. *Australian Archaeology*, 75, pp. 95–103. ISSN 03122417. DOI: 10.1080/03122417.2012.11681957.

Glyph & Cog LLC (1996). pdftotext. https://www.xpdfreader.com/pdftotext-man.html.

Golub, K., Hagelbäck, J. & Ardö, A. (2020). Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches. *Journal of Data and Information Science*, 5(1), pp. 18–38. ISSN 2096157X. DOI: 10.2478/jdis-2020-0003.

Gormley, C. & Tong, Z. (2015). *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. O'Reilly Media, Sebastopol.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O. & Quintard, L. (2011). Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. In S. Pradhan, K. Tomanek, N. Ide & A. Meyers, eds., *Proceedings of the 5th Linguistic Annotation Workshop*, pp. 92–100. Association for Computational Linguistics, Portland, Oregon, USA. https://aclanthology.org/W11-0411.

Gruber, T.R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human - Computer Studies*, 43(5-6), pp. 907–928. ISSN 10959300. DOI: 10.1006/ijhc.1995.1081.

Guo, J., Xu, G., Cheng, X. & Li, H. (2009). Named entity recognition in query. In *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, pp. 267–274. Association for Computing Machinery, Boston, Massachusetts. ISBN 9781605584836. DOI: 10.1145/1571941.1571989.

Habermehl, D. (2019). *Over zaaien en oogsten, de kwaliteit en bruikbaarheid van archeologische rapporten voor synthetiserend onderzoek*. Technical report, Rijksdienst voor Cultureel Erfgoed, Amersfoort. https://www.cultureelerfgoed.nl/publicaties/publicaties/2019/01/01/over-zaaien-en-oogsten.

Hakala, K. & Pyysalo, S. (2019). Biomedical Named Entity Recognition with Multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pp. 56–61. Association for Computational Linguistics, Hong Kong, China. DOI: 10.18653/v1/d19-5709.

Hand, D. & Christen, P. (2018). A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3), pp. 539–547. ISSN 1573-1375. DOI: 10.1007/s11222-017-9746-6.

Harris, Z.S. (1954). Distributional structure. *Word*, 10(2-3), pp. 146–162.

Hendriks, J. (2013). *Een merovingisch grafveld in het Lentseveld te Nijmegen-Noord. Evaluatie en selectierapport NLa14*. Technical report, Archeologie Gemeente Nijmegen.

Hermjakob, U., Hovy, E. & Lin, C. (2000). Knowledge-based question answering. In *Proceedings of the Sixth World Multiconference on Systems, Cybernetics, and Informatics (SCI-2002)*. International Institute of Informatics and Systemics, Winter Garden, FL.

Hessing, W., Waugh, K., van Heeringen, R. & Visser, C. (2013). *Evaluatie en optimalisatie waarderingssystematiek Kwaliteitsnorm Nederlandse Archeologie. Fase 1: Evaluatie*. Technical report, Vestigia, Amersfoort.

Highsmith, J., Paulk, M.C., Manzo, J., Mcmahon, P.E., Bowers, P., Sleve, G. & Bingham, K. (2002). Agile software development. *The journal of defense software engineering*, 15(10). http://agilesweden.com/doc/oct02.pdf.

Hinostroza, J.E., Ibieta, A., Labbé, C. & Soto, M.T. (2018). Browsing the internet to solve information problems: A study of students' search actions and behaviours using a 'think aloud' protocol. *Education and Information Technologies*, 23(5), pp. 1933–1953. DOI: 10.1007/s10639-018-9698-2.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E. & Coates, C.M. (2015). Trading consequences: A case study of combining text mining and visualization to facilitate document exploration. *Digital Scholarship in the Humanities*, 30, pp. 50–75. DOI: 10.1093/llc/fqv046.

Hjørland, B. (1997). *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Praeger. ISBN 0313298939.

Hombert, P. (1950). Les sépultures mérovingiennes par incinération en Belgique. *Revue Archéologique*, 36, pp. 96–102.

Honnibal, M. & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear, still as of February 2020*.

Hripcsak, G. & Rothschild, A.S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3), pp. 296–298. ISSN 10675027. DOI: 10.1197/jamia.M1733.

Hu, X. (2018). Usability Evaluation of E-Dunhuang Cultural Heritage Digital Library. *Data and Information Management*, 2(2), pp. 57–69. DOI: https://doi.org/10.2478/dim-2018-0008.

Huggett, J. (2012). Core or periphery? Digital humanities from an archaeological perspective. *Historical Social Research / Historische Sozialforschung*, 37(3), pp. 86–105. ISSN 01726404. DOI: 10.12759/hsr.37.2012.3.86-105.

Huurdeman, H.C. & Piccoli, C. (2020). "More than just a Picture" - The importance of context in search user interfaces for three-dimensional content. In *CHIIR 2020 - Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp. 338–342. Association for Computing Machinery, Inc, New York, NY, USA. ISBN 9781450368926. DOI: 10.1145/3343413.3377994.

Huvila, I. (2017). Being FAIR when archaeological information is MEAN: Miscellaneous, Exceptional, Arbitrary, Nonconformist. http://www.istohuvila.se/node/526.

International Committee for Documentation (CIDOC) (2014). *Information and documentation - A reference ontology for the interchange of cultural heritage information (ISO Standard No. 21127:2014)*. Technical report, International Organization for Standardization. https://www.iso.org/standard/57832.html.

Jackson, S., Richissin, C.E., McCabe, E.E. & Lee, J.J. (2020). Data-Informed Tools for Archaeological Reflexivity: Examining the substance of bone through a meta-analysis of academic texts. *Internet Archaeology*, 55. ISSN 1363-5387. DOI: 10.11141/ia.55.12.

James, E. (1988). *The Franks*. Blackwells, Oxford.

Janssens, P. & Roosens, H. (1963). Lijkverbranding en lijkbegraving op het merovingisch grafveld te Grobbendonck. *Archaeologia Belgica*, 71, pp. 265–272.

Jeffrey, S., Richards, J., Ciravegna, F., Waller, S., Chapman, S. & Zhang, Z. (2009). The Archaeotools project: faceted classification and natural language processing in an archaeological context. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367(1897), pp. 2507–19. DOI: 10.1098/rsta.2009.0038.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pp. 137–142. Springer, Berlin, Heidelberg. ISBN 978-3-540-69781-7.

Karoulis, A., Sylaiou, S. & White, M. (2006). Usability evaluation of a virtual museum interface. *Informatica*, 17(3), pp. 363–380. ISSN 08684952. DOI: 10.15388/informatica.2006.143.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. & Yih, W.t. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.550.

Khattab, O. & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48. ACM, New York, NY, USA. ISBN 9781450380164. DOI: 10.1145/3397271.3401075.

Kim, Y.M. & Lee, T.H. (2020). Korean clinical entity recognition from diagnosis text using BERT. *BMC Medical Informatics and Decision Making*, 20(S7), p. 242. ISSN 14726947. DOI: 10.1186/s12911-020-01241-8.

Kintigh, K.W. (2015). Extracting Information from Archaeological Texts. *Open Archaeology*, 1(1), pp. 96–101. DOI: 10.1515/opar-2015-0004.

Kirkpatrick, L.C. (2018). *Using Computer Screen Recordings and Think Aloud Protocols to Study Students' Cognitive Strategies While Working Online*. SAGE Publications Ltd, London. DOI: 10.4135/9781526444240.

Kleppe, M., Hendrickx, I., Veldhoen, S., Brandsen, A., Vos, H.D., Goes, K., Huang, L., Huurdeman, H., Kim, A., Mesbah, S., Reuver, M., Wang, S. & Zijdeman, R. (2019). *(Semi-) Automatic Cataloguing of Textual Cultural Heritage Objects*. Technical report, KB (National Library of the Netherlands), Den Haag. http://www.kbresearch.nl/brinkeys/report.pdf.

Koolen, M., van Gorp, J. & van Ossenbruggen, J. (2018). Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*, 34(2), pp. 368–385. ISSN 2055-7671. DOI: 10.1093/llc/fqy048.

Kudo, T. & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pp. 66–71. Association for

Computational Linguistics, Brussels, Belgium. ISBN 9781948087858. DOI: 10.18653/v1/d18-2012.

Kuratov, Y. & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. http://arxiv.org/abs/1905.07213.

Lafferty, J., Mccallum, A., Pereira, F.C.N. & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In C.E. Brodley & D.A. Pohoreckyj, eds., *Proc. 18th International Conf. on Machine Learning*, pp. 282–289. Morgan Kaufmann Publishers Inc., San Fransisco.

Lambers, K., Verschoof-van der Vaart, W. & Bourgeois, Q. (2019). Integrating Remote Sensing, Machine Learning, and Citizen Science in Dutch Archaeological Prospection. *Remote Sensing*, 11(7), p. 794. ISSN 2072-4292. DOI: 10.3390/rs11070794.

Lancaster, F.W. & Gallup, E. (1973). *Information Retrieval On-Line*. Melville Publishing Company, Los Angeles, California.

Laza, R., Pavón, R., Reboiro-Jato, M. & Fdez-Riverola, F. (2011). Evaluating the effect of unbalanced data in biomedical document classification. *Journal of integrative bioinformatics*, 8(3), p. 177. ISSN 16134516. DOI: 10.1515/jib-2011-177.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), pp. 1234–1240. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btz682.

Lehnemann, E. (2008). Das Gräberfeld von Lünen-Wethmar, Kr. Unna. *Internationale Archäologie*, 108.

Levy, T. (2014). Front Matter. *Near Eastern Archaeology*, 77(3). DOI: 10.5615/neareastarch.77.3.fm.

Lewis, C. (1982). *Using the 'thinking-aloud' method in cognitive interface design*. Technical report, IBM TJ Watson Research Center, New York.

Li, X., Zhang, H. & Zhou, X.H. (2020). Chinese clinical named entity recognition with variant neural structures based on BERT methods. *Journal of Biomedical Informatics*, 107, p. 103422. ISSN 15320464. DOI: 10.1016/j.jbi.2020.103422.

Lippok, F. (2019). Een vroegmiddeleeuws graf aan de rand van de nederzetting. In E. Norde, ed., *Nederzettingsresten uit de vroege middeleeuwen in het plangebied Leeuwesteyn Noord in Leidsche Rijn, Diemen. RAAP-rapport 3855*, pp. 91–99. RAAP.

Lippok, F. (2020). The pyre and the grave: early medieval cremation burials in the Netherlands, the German Rhineland and Belgium. *World Archaeology*, 52(1), pp. 147–162. ISSN 14701375. DOI: 10.1080/00438243.2020.1769297.

Lippok, F. (2021). The early medieval graves of Oegstgeest. In J. De Bruin, C. Bakels & F. Theuws, eds., *Oegstgeest. A riverrine settlement in the early medieval world system*, pp. 84–107. Habelt Verlag, Bonn.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv*. https://arxiv.org/abs/1907.11692.

Manning, C.D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. ISBN 0521865719.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. & Gómez-Berbís, J.M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards and Interfaces*, 35(5), pp. 482–489. ISSN 09205489. DOI: 10.1016/j.csi.2012.09.004.

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McHugh, M.L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), pp. 276–282. ISSN 13300962. DOI: 10.11613/bm.2012.031.

Meckseper, C. & Warwick, C. (2003). The Publication of Archaeological Excavation Reports Using XML. *Literary and Linguistic Computing*, 18(1), pp. 63–75. ISSN 0268-1145. DOI: 10.1093/llc/18.1.63.

Mélanie-becquet, F., Ferguth, J., Gruel, K. & Poibeau, T. (2015). Archaeology in the Digital Age: From Paper to Databases. In *Proceedings of the conference "Digital Humanities 2015"*. Sydney.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Ministerie van Onderwijs Cultuur en Wetenschap (2007). Wet op de archeologische monumentenzorg. https://wetten.overheid.nl/jci1.3:c:BWBR0021162&z=2008-01-01&g=2008-01-01.

Ministerie van Onderwijs Cultuur en Wetenschap (2015). Erfgoedwet. https://wetten.overheid.nl/jci1.3:c:BWBR0037521&z=2021-07-01&g=2021-07-01.

Mohammed, R., Rawashdeh, J. & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pp. 243–248. DOI: 10.1109/ICICS49469.2020.239556.

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2013). *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2nd edition. ISBN 9780262039406.

Moon, T., Awasthy, P., Ni, J. & Florian, R. (2019). Towards Lingua Franca Named Entity Recognition with BERT. *arXiv*. http://arxiv.org/abs/1912.01389.

Morgan, C. & Eve, S. (2012). DIY and digital archaeology: What are you doing to participate? *World Archaeology*, 44(4), pp. 521–537. ISSN 00438243. DOI: 10.1080/00438243.2012.741810.

Nadkarni, P.M., Ohno-Machado, L. & Chapman, W.W. (2011). Natural language processing: An introduction. DOI: 10.1136/amiajnl-2011-000464.

Nakayama, H. (2019). chakki-works/doccano: Open source text annotation tool for machine learning practitioner. https://github.com/chakki-works/doccano.

Niccolucci, F. & Richards, J. (2019). *The ARIADNE Impact*. Archaeolingua Foundation, Hungary. DOI: 10.5281/ZENODO.4319058.

Niccolucci, F. & Richards, J.D. (2013). ARIADNE: Advanced Research Infrastructures for Archaeological Dataset Networking in Europe. *International Journal of Humanities and Arts Computing*, 7(1-2), pp. 70–88. ISSN 1753-8548. DOI: 10.3366/ijhac.2013.0082.

Nielsen, J. & Landauer, T.K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93*, pp. 206–213. ACM Press, New York, New York, USA. ISBN 0897915755. DOI: 10.1145/169059.169166.

Norvig, P. (2013). English Letter Frequency Counts: Mayzner Revisited. http://norvig.com/mayzner.html.

Nozza, D., Bianchi, F. & Hovy, D. (2020). What the [MASK]? Making Sense of Language-Specific BERT Models. *arXiv*. ISSN 23318422. http://arxiv.org/abs/2003.02912.

Paijmans, H. & Brandsen, A. (2009). What is in a Name: Recognizing Monument Names from Free-Text Monument Descriptions. In M. van Erp, J. Stehouwer & M. van Zaanen, eds., *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning (Benelearn)*, pp. 2–6. Tilburg centre for Creative Computing, Tilburg. http://benelearn09.uvt.nl/Proceedings_Benelearn_09.pdf.

Paijmans, H. & Brandsen, A. (2010). Searching in archaeological texts: Problems and solutions using an artificial intelligence approach. *PalArch's Journal Of Archaeology Of Egypt/Egyptology*, 7(2), pp. 1–6.

Paijmans, J. & Wubben, H. (2008). Preparing archeological reports for intelligent retrieval. In A. Posluschny, K. Lambers & I. Herzog, eds., *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA)*, pp. 2–6. Berlin.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct), pp. 2825–2830.

Pescarin, S., Pagano, A., Wallergård, M., Hupperetz, W. & Ray, C. (2014). Evaluating Virtual Museums: Archeovirtual Case Study. In P. Verhagen & G. Earl, eds., *Archaeology in the Digital Era*, pp. 74–82. Amsterdam University Press. DOI: doi:10.1515/9789048519590-009.

Peters, T. (2004). PEP 20 – The Zen of Python. https://www.python.org/dev/peps/pep-0020/.

Peterson, C. & Seligman, M. (1984). *Content analysis of verbatim explanations: The CAVE technique for assessing explanatory style*. Technical report, Virginia Polytechnic Institute and State University.

Plets, G., Huijnen, P. & van Oeveren, D. (2021). Excavating Archaeological Texts: Applying Digital Humanities to the Study of Archaeological Thought and Banal Nationalism. *Journal of Field Archaeology*, pp. 1–14. ISSN 0093-4690. DOI: 10.1080/00934690.2021.1899889.

Postma, M., van Miltenburg, E., Segers, R., Schoen, A. & Vossen, P. (2016). Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*, pp. 300–308. Bucharest, Romania.

Powers, D. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2(1), pp. 37–63.

Pressman, R.S. (2005). *Software engineering : a practitioner's approach*. McGraw-Hill College, New York, NY, USA, 6th edition. ISBN 007301933X.

Rabinowitz, A., Shaw, R., Buchanan, S., Golden, P. & Kansa, E. (2016). Making Sense of the Ways we make Sense of the Past: The PeriodO Project. *Bulletin of the Institute of Classical Studies*, 59(2), pp. 42–55. ISSN 0076-0730. DOI: 10.1111/j.2041-5370.2016.12037.x.

Ramshaw, L.A. & Marcus, M.P. (1999). Text Chunking Using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*, pp. 157–176. Association for Computational Linguistics. DOI: 10.1007/978-94-017-2390-9_10.

Rau, L.F. (1991). Extracting company names from text. In *Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications*, pp. 29–32. Publ by IEEE, Miami Beach, FL, USA. ISBN 0818621354. DOI: 10.1109/caia.1991.120841.

Renfrew, C. & Bahn, P.G. (2019). *Archaeology: theories, methods and practice (8th edition)*. Thames and Hudson London.

Richards, J., Tudhope, D. & Vlachidis, A. (2015). Text Mining in Archaeology: Extracting Information from Archaeological Reports. In J.A. Barcelo & I. Bogdanovic, eds., *Mathematics and Archaeology*, pp. 240–254. CRC Press, Boca Raton. DOI: 10.1201/b18530-15.

Rico, M., Vila-Suero, D., Botezan, I. & Gómez-Pérez, A. (2019). Evaluating the impact of semantic technologies on bibliographic systems: A user-centred and comparative approach. *Journal of Web Semantics*, 59. DOI: 10.1016/J.WEBSEM.2019.03.001.

Rieh, S. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, 42(3), pp. 751–768. DOI: doi.org/10.1016/j.ipm.2005.05.005.

Rijksdienst voor het Cultureel Erfgoed (2019a). Archeologisch onderzoek - aantal onderzoeksmeldingen | Erfgoedmonitor. https://erfgoedmonitor.nl/indicatoren/archeologisch-onderzoek-aantal-onderzoeksmeldingen.

Rijksdienst voor het Cultureel Erfgoed (2019b). Archis. https://archis.cultureelerfgoed.nl.

Rosenzweig, E. (2015). *Successful user experience: Strategies and roadmaps*. Elsevier, Waltham, MA, USA. ISBN 9780128010617. DOI: 10.1016/c2013-0-19353-1.

Roth, B.J. (2010). An Academic Perspective on Grey Literature. *Archaeologies*, 6(2), pp. 337–345. ISSN 1555-8622. DOI: 10.1007/s11759-010-9141-9.

Rural Riches project (2021). The Rural Riches database. https://www.merovingianarchaeology.org/blog/about/the-rr-database.

Russell-Rose, T., Chamberlain, J. & Azzopardi, L. (2018). Information retrieval in the workplace: A comparison of professional search practices. *Information Processing and Management*, 54(6), pp. 1042–1057. ISSN 03064573. DOI: 10.1016/j.ipm.2018.07.003.

Russell-Rose, T. & Shokraneh, F. (2020). Designing the Structured Search Experience: Rethinking the Query-Builder Paradigm. *Weave: Journal of Library User Experience*, 3(1). ISSN 2333-3316. DOI: 10.3998/weave.12535642.0003.102.

Salton, G. (1968). *Automatic Information Organization and Retrieval*. McGraw Hill Text. ISBN 0070544859.

Salton, G. (1971). *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, Upper Saddle River, NJ.

Sasaki, Y. (2007). *The truth of the F-measure*. Technical report, School of Computer Science, University of Manchester, Manchester. https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.

Schmidt, S.C. & Marwick, B. (2020). Tool-Driven Revolutions in Archaeological Science. *Journal of Computer Applications in Archaeology*, 3(1), pp. 18–32. ISSN 2514-8362. DOI: 10.5334/jcaa.29.

Selhofer, H. & Geser, G. (2014). *D2.1: First Report on Users' Needs*. Technical report, ARIADNE. http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/07/ARIADNE_D2-1_First_report_on_users_needs.pdf.

Seok, M., Song, H.J., Park, C.Y., Kim, J.D. & Kim, Y.S. (2016). Named entity recognition using word embedding as a feature. *International Journal of Software Engineering and its Applications*, 10, pp. 93 – 104. ISSN 17389984. DOI: 10.14257/ijseia.2016.10.2.08.

Seymour, D.J. (2010). Sanctioned Inequity and Accessibility Issues in the Grey Literature in the United States. *Archaeologies*, 6(2), pp. 233–269. ISSN 1555-8622. DOI: 10.1007/s11759-010-9144-6.

Sienčnik, S.K. (2015). Adapting word2vec to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015*, 109, pp. 239–243. Linköping University Electronic Press, Vilnius, Lithuania.

Song, Y., Zhou, D. & He, L.W. (2011). Post-ranking query suggestion by diversifying search results. In *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 815–824. Association for Computing Machinery, Beijing, China. ISBN 9781450309349. DOI: 10.1145/2009916.2010025.

Sorel, D. (2018). jQuery QueryBuilder. https://querybuilder.js.org/.

Soto, A., Olivas, J.A. & Prieto, M.E. (2008). Fuzzy approach of synonymy and polysemy for information retrieval. *Studies in Fuzziness and Soft Computing*, 224, pp. 179–198. ISSN 14349922. DOI: 10.1007/978-3-540-76973-6_12.

Souza, F., Nogueira, R. & Lotufo, R. (2019). Portuguese Named Entity Recognition using BERT-CRF. *arXiv*. ISSN 23318422. http://arxiv.org/abs/1909.10649.

Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing & Management*, 38(3), pp. 401–426. DOI: 10.1016/S0306-4573(01)00036-X.

Sporleder, C. (2010). Natural Language Processing for Cultural Heritage Domains. *Language and Linguistics Compass*, 4(9), pp. 750–768. ISSN 1749818X. DOI: 10.1111/j.1749-818X.2010.00230.x.

Steiner, C.M., Agosti, M., Sweetnam, M.S., Hillemann, E.C., Orio, N., Ponchia, C., Hampson, C., Munnelly, G., Nussbaumer, A., Albert, D. & Conlan, O. (2014). Evaluating a digital humanities research environment: the CULTURA approach. *International Journal on Digital Libraries*, 15(1), pp. 53–70. DOI: 10.1007/s00799-014-0127-x.

Stichting Infrastructuur Kwaliteitsborging Bodembeheer (2016). BRL 4000. https://www.sikb.nl/archeologie/richtlijnen/brl-4000.

Strubell, E., Ganesh, A. & McCallum, A. (2020). Energy and policy considerations for deep learning in NLP. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 3645–3650. Association for Computational Linguistics, Florence, Italy. ISBN 9781950737482. DOI: 10.18653/v1/p19-1355.

Talboom, L. (2017). *Improving the discoverability of zooarchaeological data with the help of Natural Language Processing.* Thesis, University of York.

Talja, S. (1997). Constituting "information" and "user" as research objects: A theory of knowledge formations as an alternative to the information man-theory. *Information seeking in context*, pp. 67–80.

Talks, A. (2019). *An exploration of NLP and NER for enhanced search in osteoarchaeological and palaeopathological textual resources.* Thesis, University of York.

Tanasi, D. (2020). The digital (within) archaeology. Analysis of a phenomenon. *The Historian*, 82(1), pp. 22–36. ISSN 0018-2370. DOI: 10.1080/00182370.2020.1723968.

Theunissen, L. & Feiken, R. (2014). *Analyse archeologische kenniswinst (2000 - 2014).* Technical report, Rijksdienst voor het Cultureel Erfgoed, Amersfoort.

Thomsett-Scott, B.C. (2006). Web site usability with remote users: Formal usability studies and focus groups. *Journal of Library Administration*, 45(3-4), pp. 517–547. ISSN 01930826. DOI: 10.1300/J111v45n03_14.

Tikhomirov, M., Loukachevitch, N., Sirotina, A. & Dobrov, B. (2020). Using bert and augmentation in named entity recognition for cybersecurity domain. In *Natural Language Processing and Information Systems*, volume 12089 LNCS, pp. 16–24. Springer International Publishing, Cham. ISBN 9783030513092. ISSN 16113349. DOI: 10.1007/978-3-030-51310-8_2.

Tjong Kim Sang, E.F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Tong, Z. (2018). elasticsearch-php. https://github.com/elastic/elasticsearch-php.

Traviglia, A. & Torsello, A. (2017). Landscape Pattern Detection in Archaeological Remote Sensing. *Geosciences*, 7(4), p. 128. ISSN 2076-3263. DOI: 10.3390/geosciences7040128.

Trier, Ø.D., Salberg, A.B. & Pilø, L.H. (2018). Semi-automatic mapping of charcoal kilns from airborne laser scanning data using deep learning. In *CAA2016: Oceans of Data. Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology*, pp. 219–231. Archaeopress Oxford.

Truyens, M. & Van Eecke, P. (2014). Legal aspects of text mining. *Computer Law & Security Review*. ISSN 02673649. DOI: 10.1016/j.clsr.2014.01.009.

Tudhope, D., May, K., Binding, C. & Vlachidis, A. (2011). Connecting archaeological data and grey literature via semantic cross search. *Internet archaeology*, 30. DOI: doi.org/10.11141/ia.30.5.

Tunkelang, D. (2009). Faceted Search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1), pp. 1–80. ISSN 1947-945X. DOI: 10.2200/s00190ed1v01y200904icr005.

Van den Bosch, A., Busser, B., Canisius, S. & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman & V. Vandeghinste, eds., *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pp. 99–114. Leuven.

Van den Dries, M. (2016). Is everybody happy? User satisfaction after ten years of quality management in European archaeological heritage management. In P. Florjanowicz, ed., *When Valletta meets Faro, the reality of European archaeology in the 21st century, proceedings of the International Conference*, pp. 126–135. Archaeolingua, Lisbon.

Van Es, W. (1968). *Grafritueel en Kerstening*. Inaugural lecture Vrije Universiteit Amsterdam, Amsterdam.

Van Es, W. & Schoen, R. (2008). Het vroegmiddeleeuwse grafveld van Zweeloo. *Palaeohistoria*, 45/50, pp. 795–935. https://ugp.rug.nl/Palaeohistoria/article/view/25161.

Van Gompel, M. & Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3, pp. 63–81.

Van Haperen, M. (2017). *Early Medieval Grave Reopenings in the Low Countries*. Thesis, Leiden University. DOI: 10.17026/dans-x6b-bvgj.

Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T. & Van de Walle, R. (2015). Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*, 30(2), pp. 262–279. ISSN 2055-7671. DOI: 10.1093/llc/fqt067.

Van Waes, L. (2000). Thinking aloud as a method for testing the usability of Websites: the influence of task variation on the evaluation of hypertext. *IEEE Transactions on Professional Communication*, 43(3), pp. 279–291. DOI: 10.1109/47.867944.

Van Zundert, J.J. (2016). The case of the bold button: Social shaping of technology and the digital scholarly edition. *Digital Scholarship in the Humanities*, 31(4), pp. 898–910. ISSN 2055-7671. DOI: 10.1093/llc/fqw012.

Verberne, S., Boves, L. & Bosch, A. (2016). Information access in the art history domain. Evaluating a federated search engine for Rembrandt research. *Digital Humanities Quarterly*, 10(4), p. online. ISSN 1938-4122. http://www.digitalhumanities.org/dhq/vol/10/4/000265/000265.html.

Verberne, S., He, J., Kruschwitz, U., Wiggers, G., Larsen, B., Russell-Rose, T. & de Vries, A.P. (2019). First International Workshop on Professional Search. *ACM SIGIR Forum*, 52(2), pp. 153–162. ISSN 0163-5840. DOI: 10.1145/3308774.3308799.

Verschoof-van der Vaart, W. & Brandsen, A. (2020). Boundingbox Localizer Tool (BLT) - Brandenburgische Technische Universität Cottbus-Senftenberg version. *Zenodo Repository*. DOI: 10.5281/ZENODO.3888053.

Verschoof-Van Der Vaart, W.B., Lambers, K., Kowalczyk, W. & Bourgeois, Q.P. (2020). Combining deep learning and location-based ranking for large-scale archaeological prospection of LiDAR data from the Netherlands. *ISPRS International Journal of Geo-Information*, 9(5), p. 293. ISSN 22209964. DOI: 10.3390/ijgi9050293.

Verschoof-van der Vaart, W.B. & Landauer, J. (2021). Using CarcassonNet to automatically detect and trace hollow roads in LiDAR data from the Netherlands. *Journal of Cultural Heritage*, 47, pp. 143–154. ISSN 12962074. DOI: 10.1016/j.culher.2020.10.009.

Verwers, W. & van Tent, W. (2015). *Merovingisch grafveld Elst-'t Woud. Rapportage Archeologische Monumentenzorg 223*. Technical report, Rijksdienst voor het Cultureel Erfgoed, Amersfoort.

Vince, A. (1996). Editorial. *Internet Archaeology*, 1. ISSN 13635387. DOI: 10.11141/ia.1.7.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F. & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv*. http://arxiv.org/abs/1912.07076.

Vlachidis, A. (2012). Semantic Indexing via Knowledge Organization Systems: Applying the CIDOC-CRM to Archaeological Grey Literature. *Unpublished PhD Thesis, University of South Wales (USW)*.

Vlachidis, A., Binding, C., May, K. & Tudhope, D. (2013). Automatic metadata generation in an archaeological digital library: Semantic annotation of grey literature. *Studies in Computational Intelligence*, 458, pp. 187–202. ISSN 1860949X. DOI: 10.1007/978-3-642-34399-5_10.

Vlachidis, A. & Tudhope, D. (2012). A pilot investigation of information extraction in the semantic annotation of archaeological reports. *International Journal of Metadata, Semantics and Ontologies*, 7(3), p. 222. ISSN 1744-2621. DOI: 10.1504/IJMSO.2012.050183.

Vlachidis, A. & Tudhope, D. (2016). A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. *Journal of the Association for Information Science and Technology*, 67(5), pp. 1138–1152. DOI: 10.1002/asi.23485.

Vlachidis, A., Tudhope, D., Wansleeben, M., Azzopardi, J., Green, K., Xia, L. & Wright, H. (2017). *D16.4: Final Report on Natural Language Processing.* Technical report, ARIADNE. http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/01/D16.4_Final_Report_on_Natural_Language_Processing_Final.pdf.

Voorhees, E. (2001). Overview of TREC 2001. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, pp. 1–13.

Wamers, E. (2015). *Das bi-rituelle Kinderdoppelgrab der späten Merowingerzeit unter der Frankfurter Bartholomäuskirche (»Dom«).* Schnell and Steiner, Regensburg.

Wei, J. & Zou, K. (2020). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 6382–6388. ISBN 9781950737901. DOI: 10.18653/v1/d19-1670.

Wen, M., Vasthimal, D.K., Lu, A., Wang, T. & Guo, A. (2019). Building large-scale deep learning system for entity recognition in e-commerce search. In *BDCAT 2019 - Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, pp. 149–154. Association for Computing Machinery, Inc, New York, New York, USA. ISBN 9781450370165. DOI: 10.1145/3365109.3368765.

Wesson & Cottier (2014). Big Sites, Big Questions, Big Data, Big Problems: Scales of Investigation and Changing Perceptions of Archaeological Practice in the Southeastern United States. *Bulletin of the History of Archaeology*, 24(0), p. 16. ISSN 2047-6930. DOI: 10.5334/bha.2416.

Wheatley, D. (2004). Making space for an archaeology of place. *Internet Archaeology*, 15. DOI: 10.11141/ia.15.10.

Wilcke, W.X., de Boer, V., de Kleijn, M.T., van Harmelen, F.A. & Scholten, H.J. (2019). User-centric pattern mining on knowledge graphs: An archaeological case study. *Journal of Web Semantics*, 59, pp. 1–10. ISSN 15708268. DOI: 10.1016/j.websem.2018.12.004.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E.,

Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3. ISSN 2052-4463. DOI: 10.1038/sdata.2016.18.

Williams, H. (2014). A well-urned rest: Cremation and inhumation in early Anglo-Saxon England. In I. Kuijt, C.P. Quinn & G. Cooney, eds., *Transformation by fire: The archaeology of cremation in cultural context*, pp. 93–118. University of Arizona Press, Tucson.

Wiseman, R. & Ronn, P. (2020). *Archaeology on Furlough: Accessing Archaeological Information Online: A Survey of Volunteers' Experiences*. Technical report, Cambridge University. DOI: 10.17863/CAM.54876.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, Stroudsburg, PA, USA. DOI: 10.18653/v1/2020.emnlp-demos.6.

Wu, S. & Dredze, M. (2020). Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130. Association for Computational Linguistics, Stroudsburg, PA, USA. DOI: 10.18653/v1/2020.repl4nlp-1.16.

Xiong, Y., Huang, Y., Chen, Q., Wang, X., Ni, Y. & Tang, B. (2020). A joint model for medical named entity recognition and normalization. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2664, pp. 499–504. ISSN 16130073.

Yamada, I., Asai, A., Shindo, H., Takeda, H. & Matsumoto, Y. (2020). LUKE: Deep Contextualized Entity Representations with Entity-aware Self-

attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6442–6454. Association for Computational Linguistics, Stroudsburg, PA, USA. DOI: 10.18653/v1/2020.emnlp-main.523.

Zubrow, E.B. (2006). Digital archaeology: A historical context. In P. Daly & T. Evans, eds., *Digital Archaeology: Bridging Method and Theory*. Routledge, London, 1st edition. ISBN 9780415310505. DOI: 10.4324/9780203005262.

Appendices

A

Category frequencies

Site type categories frequency overview									
Label	Freq	Label	Freq	Label	Freq	Label	Freq	Label	Freq
xxx	33532								
cthd	1463	bewv.wp	10	**idnh**	2015	sv.vorg	0	bgv.meg	9
cthd.x	379	bewv.n	0	idnh.x	1074	sv.bsb	0	**bgr**	1502
cthd.klo	299	bewv.rv	501	idnh.tk	0	**gw**	169	bgr.gvic	1502
cthd.kpl	24	bewv.stel	4	idnh.tn	128	gw.x	40	**infr**	7327
cthd.sgmw	0	bewv.bw	0	idnh.br	36	gw.vw	58	infr.x	1248
cthd.kerk	581	bewv.hp	1566	idnh.zp	0	gw.hout	0	infr.weg	1575
cthd.rcp	372	bewv.th	0	idnh.sb	71	gw.ijw	9	infr.dam	53
cthd.oloc	1	bewv.inka	0	idnh.hkb	127	gw.zw	3	infr.werf	0
cthd.temp	2	bewv.sv	0	idnh.bb	5	gw.kw	50	infr.gem	5
bewv	25264	bewv.bext	5872	idnh.ll	112	gw.griw	0	infr.rede	0
bewv.x	15236	bewv.vkm	63	idnh.hb	4	gw.mw	4	infr.per	3766
bewv.lg	89	bewv.tw	388	idnh.m	150	gw.vsw	8	infr.strek	0
bewv.wb	51	bewv.lw	130	idnh.rom	1	**bgv**	6317	infr.wat	228
bewv.sch	65	**apvv**	3152	idnh.wam	2	bgv.x	731	infr.dui	221
bewv.vx	815	apvv.x	1415	idnh.wim	1	bgv.gvc	522	infr.vijv	0
bewv.vlp	102	apvv.vw	0	idnh.gp	0	bgv.tpgb	0	infr.kan	273
bewv.lk	0	apvv.vk	166	idnh.pb	227	bgv.gvi	536	infr.slu	103
bewv.ct	0	apvv.vs	6	idnh.vb	312	bgv.gvx	983	infr.kslu	0
bewv.cstl	5	apvv.stel	0	idnh.mb	388	bgv.kh	605	infr.lv	0
bewv.mbh	125	apvv.ek	0	idnh.mbnf	0	bgv.rgv	1	infr.hav	982
bewvv.pls	0	apvv.cf	23	idnh.mbf	0	bgv.ghv	2476	infr.kade	1
bewv.kwb	490	apvv.dp	142	idnh.kb	2	bgv.bhv	0	infr.vweg	7
bewv.ht	332	apvv.la	1879	**sv**	971	bgv.vgv	0	infr.brug	256
bewv.aw	7	apvv.ak	0	sv.x	971	bgv.cjbp	536	infr.dok	0
bewv.dump	0	apvv.tuin	20	sv.obsb	0	bgv.uv	1295	infr.vs	0
bewv.vic	332	apvv.pdek	2	sv.ijz	0	bgv.gh	1221	infr.vrde	1
bewv.kaze	0	**wrak**	384	sv.h	0	bgv.gx	731	infr.spre	0
bewv.fort	8	wrak.schip	384	sv.lad	0	bgv.vg	163	infr.watw	11
bewv.sk	2470	wrak.vlgtg	5	sv.hijz	0	bgv.dier	131	infr.dij	721

Table A.1: An overview of the frequencies for all site type categories. Main categories are denoted in bold.

Time periods categories frequency overview									
Label	Freq	Label	Freq	Label	Freq	Label	Freq	Label	Freq
paleo	2077	neov	6459	bronsm	8494	romvb	12414	vmec	11767
paleov	1197	neova	6456	bronsma	8397	romm	12427	vmed	12194
paleom	1460	neovb	6445	bronsmb	8312	romma	12381	lme	18832
paleol	1816	neom	6127	bronsl	7910	rommb	12348	lmea	17053
paleola	1732	neoma	6098	**ijz**	13876	roml	11939	lmeb	18235
paleolb	1816	neomb	5893	ijzv	10356	romla	11921	**nt**	19833
meso	4290	neol	8954	ijzm	11307	romlb	11850	nta	17511
mesov	3133	neola	8200	ijzl	12033	**xme**	20593	ntb	17514
mesom	3152	neolb	8947	**rom**	13299	vme	12642	ntc	18525
mesol	4180	**brons**	10414	romv	12421	vmea	11645		
neo	9916	bronsv	8380	romva	12275	vmeb	11874		

Table A.2: An overview of the frequencies for all time period categories. Main categories are denoted in bold.

B

Filter list

Terms used for document filtering	
List name	**Terms**
genList	notulen, bijlage, meta
rapList	dagrapport, dag_rapport, weekrapport, week_rapport, weekverslag, week_verslag, logboek
pvaList	draaiboek, plan_van_aanpak, pva
omnList	onderzoeksmeldingsnummer, onderzoeksmeldings_nummer, onderzoeks_meldings_nummer
totList	rapList + pvaList + pveList + omnList + genList

Table B.1: An overview of different types of lists and included terms.

C

Category frequencies test set

Time periods categories frequency overview test set									
Label	Freq	Label	Freq	Label	Freq	Label	Freq	Label	Freq
paleo	13	neov	24	bronsm	15	romvb	25	vmec	30
paleov	7	neova	24	bronsma	14	romm	27	vmed	29
paleom	8	neovb	24	bronsmb	15	romma	27	lme	48
paleol	12	neom	23	bronsl	18	rommb	27	lmea	41
paleola	12	neoma	23	**ijz**	37	roml	25	lmeb	48
paleolb	12	neomb	23	ijzv	34	romla	25	**nt**	49
meso	21	neol	26	ijzm	28	romlb	25	nta	42
mesov	17	neola	25	ijzl	30	**xme**	54	ntb	43
mesom	18	neolb	26	**rom**	29	vme	30	ntc	39
mesol	20	**brons**	24	romv	25	vmea	28		
neo	29	bronsv	19	romva	25	vmeb	28		

Table C.1: An overview of the frequencies for all time period categories captured by the reference test set. Main categories are denoted in bold.

Site type categories frequency overview test set									
Label	Freq	Label	Freq	Label	Freq	Label	Freq	Label	Freq
cthd	1	bewv.hp	7	idnh.hkb	2	bgv.x	4	**bgr**	3
cthd.klo	1	bewv.bext	3	idnh.ll	1	bgv.gvc	2	bgr.gvic	3
bewv	65	**apvv**	8	idnh.m	1	bgv.gvi	3	**infr**	19
bewv.x	53	apvv.x	3	idnh.pb	1	bgv.gvx	3	infr.x	1
bewv.vx	1	apvv.cf	1	idnh.vb	2	bgv.kh	1	infr.weg	4
bewv.vlp	1	apvv.la	4	idnh.mb	2	bgv.ghv	6	infr.per	6
bewv.kwb	1	**wrak**	2	**sv**	1	bgv.cjbp	3	infr.kan	2
bewv.ht	10	wrak.schip	2	sv.x	1	bgv.uv	4	infr.brug	2
bewv.vic	1	**idnh**	11	**gw**	1	bgv.gx	4	infr.dij	5
bewv.sk	4	idnh.x	4	gw.vw	1	bgv.vg	1	**xxx**	17
bewv.rv	1	idnh.tn	1	**bgv**	16	bgv.dier	1		

Table C.2: An overview of the F1 scores for the main and sub-categories for site type classification as captured by the reference test set. Sub-categories not present within the reference test set are not included. Again, main categories are denoted in bold.

D

Curriculum Vitae

Alex **Brandsen**

PhD Candidate · Code Wrangler

Leiden University, Faculty of Archaeology, Einsteinweg 2, 2333CC Leiden, The Netherlands

+31 681 833 764 | a.brandsen@arch.leidenuniv.nl | alexbrandsen.nl | alexbrandsen | alex-brandsen | @alex_brandsen

"Programming isn't about what you know; it's about what you can figure out." - Chris Pine

Summary

I studied archaeology for my BA and MSc, but always with a focus on digital techniques and web technologies. After my studies I pursued a career in web development, working on front and back end as well as server maintenance and client liaison for six years. I completed a PhD in Digital Archaeology at Leiden University in 2022, researching the use of text mining in archaeological grey literature. I am currently a postdoc in the EXALT project, further building on my PhD research.

I'm a fast learner, tinkerer, and passionate about open science and reproducibility.

Work Experience

Leiden University
Leiden, The Netherlands

Postdoc
June 2021 - PRESENT

- Further building on my PhD research in the EXALT project
- Creating a multilingual search engine for all available archaeological texts about the Netherlands and surrounding areas
- Developing and teaching courses at BA and MA level

Leiden University
Leiden, The Netherlands

PhD Candidate
May 2017 - May 2021

- Investigating the use of Text Mining techniques to make archaeological excavation reports more accessible
- Using machine learning techniques to perform Named Entity Recognition
- Building an intuitive search UI for archaeologists
- Teaching assistance in various courses (Databases, Text Mining, etc)

Space Creative / The Wrapped Agency
Leeds, United Kingdom

Web Developer
Jan 2010 - Apr. 2017

- Developing websites on the Magento eCommerce and Concrete5 CMS platforms, as well as designing and developing custom applications
- Worked on complex projects from initial specification and database design right up to the final stages of responsive testing
- Gained experience in Linux sysadmin & server maintenance, custom (Google) mapping, geographical searches, XML, complex jQuery applications, Photoshop, Inkscape & Illustrator

Impulse Media
York, United Kingdom

Freelance Web Developer
Sep. 2010 - Dec. 2010

- Mainly working for Impulse Media, developing PHP web applications, converting PSD files to HTML/CSS templates, developing iPhone apps and working with Concrete5 CMS and eCommerce platform Magento

BioArch, University of York
York, United Kingdom

Internship: Assistent Web Developer
Jan. 2010 - Jun. 2010

- Assisted David Harker in developing SHAARKWeb, a web-based UI that inputs and processes information from users of the NEAAR lab
- Used Object-Oriented PHP, Propel, XHTML and CSS to develop parts of the system; mainly the login system, AJAX dropdown boxes populated by external database queries and programmatically importing MS Excel spreadsheets into the database

Antiquity Journal
York, United Kingdom

Internship: Assistent Web Designer
Oct. 2009 - Dec. 2009

- Created a new Project Gallery Archive homepage

The Open Boek Project, at Rijksdienst voor Cultureel Erfgoed
Amersfoort, The Netherlands

Internship: Assistant Web/Software Developer
Nov. 2008 - Mar. 2008

- Assisted in making the smart index- and search engine 'Open Boek' more user-friendly for archaeologists
- Adapted the system to be able to process Dutch texts as well as English texts
- Gained experience with LaTeX, MySQL, PHP, AWK and Bash-shell scripting

Education

University of York — *York, United Kingdom*

MSc in Archaeological Information Systems — *Sep. 2010*

- Grade: 1st Class Distinction
- Major in Archaeological Information Systems
- Main Topics: Web Design, Database Design, Geographical Information Systems & Virtual Reality
- Minor in Zooarchaeology
- Dissertation: Digital Medieval Graffiti; Using online-GIS to display and edit non-geographical data

Leiden University — *Leiden, The Netherlands*

BA in Archaeology — *Sep. 2009*

- Grade: 7.5
- Major in European Prehistory
- Minor in Archaeological Information Systems (Database Design, GIS & Use of Total Station)
- Additional courses: Physical Anthropology
- Thesis Topic: Using Ground Penetrating Radar to assess burial mounds

Tabor College, Location Werenfridus — *Hoorn, The Netherlands*

VWO (Pre-university secondary education) — *Jul. 2005*

- Specialisation: Economics and Society. Main topics are History and Economics
- Additional course: Computer Science (MS Access, SQL, Java & HTML)

Publications

Brandsen, A, & Lippok, F, 2021. A burning question – Using an intelligent grey literature search engine to change our views on early medieval burial practices in the Netherlands. *Journal of Archaeological Science*, 133. DOI: 10.1016/j.jas.2021.105456

Brandsen, A & Koole, M, 2021. Labelling the Past: Data Set Creation and Multi-label Classification of Dutch Archaeological Excavation Reports. *Language Resources and Evaluation*. DOI: 10.1007/s10579-021-09552-6

Brandsen, A, Verberne, S, Lambers, K & Wansleeben, M, 2020. Creating a Dataset for Named Entity Recognition in the Archaeology Domain. *Proceedings of the 12th Language Resources and Evaluation Conference*, pp.4573-4577. URL: www.aclweb.org/anthology/2020.lrec-1.562

Brandsen, A, Lambers, K, Verberne, S & Wansleeben, M, 2019. User Requirement Solicitation for an Information Retrieval System Applied to Dutch Grey Literature in the Archaeology Domain. *Journal of Computer Applications in Archaeology*, 2(1), pp.21–30. DOI: 10.5334/jcaa.33

Paijmans, H & Brandsen, A, 2010. Searching in Archaeological Texts. Problems and Solutions Using an Articial Intelligence Approach, *PalArchs Journal of Archaeology of Egypt/Egyptology*, 7(2).

Paijmans, H & Brandsen, A, 2009. What is in a Name: Recognizing Monument Names from Free-Text Monument Descriptions, *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*. Tilburg centre for Creative Computing, Tilburg.

Selected Presentations

2021	**CLIN31,** Can BERT Dig It? Named Entity Recognition for Information Retrieval in the Archaeology Domain	*Online*
2020	**Keeping Archaeology Together,** Using Transfer Learning for NER in Dutch Excavation Reports	*Online*
2020	**CLIN30**, BERT-NL: a set of language models pre-trained on the Dutch SoNaR corpus	*Utrecht*
2019	**Machine Learning in Archaeology**, Using Machine Learning for NER in Dutch Excavation Reports	*Rome*
2019	**CAA**, User Interface Design and Evaluation for Online Professional Search in Dutch Archaeology	*Krakow*
2019	**ICT Open,** Brinkey-generator - Computer Assisted Assignment of Thesaurus Topics for Scientific Texts	*Hilversum*
2018	**EAA**, Utilising Text Mining to Unlock the Hidden Knowledge in Dutch Archaeological Reports	*Barcelona*
2018	**DHBenelux 2018,** Knowledge dissemination and discovery in Dutch excavation data	*Amsterdam*
2017	**DIR2017,** Archaeological Entity Recognition for Information Retrieval in Dutch Archaeological Reports	*Hilversum*

Committees

2020	**Committee member**, ARCHON Digital Archaeology Workgroup	*Amsterdam*
2020	**Outreach Officer**, CAA NL/FL	*Amsterdam*
2019	**Main organiser**, Digital Archaeology Group	*Leiden*
2019	**Committee member**, ARCHON Digital Archaeology Workgroup	*Leiden*
2018	**Committee member**, Dutch-Belgian Information Retrieval Workshop 2018	*Leiden*
2018	**Committee member**, Digital Archaeology Workshops Netherlands-Flanders (DAWN) 2018	*Leiden*

Glossary

ABR *Archeologisch Basisregister*. 33, 52–57, 62, 63, 82, 89, 124

ADS Archaeology Data Service. 30, 79

AGNES Archaeological Grey-literature Named Entity Search. 3–10, 37, 48, 53, 78, 79, 83, 90, 92–99, 110, 111, 140, 142–146, 148–151, 153, 154, 156, 158–161, 167–171, VI, X

ALICE Academic Leiden Interdisciplinary Cluster Environment. 33

API Application Programming Interface. 30, 79, 90, 125

Archis *Archeologisch Informatiesysteem*, a system for registering and accessing data about archaeological research, finds and monuments in the Netherlands, maintained by the RCE. 87, 94, 95, 141, 150, 151, 153, X

ARIADNE Advanced Research Infrastructure for Archaeological Dataset Networking in Europe. 30, 31, 34, 37, 38, 79, 81, 94, 174

BERT Bidirectional Encoder Representations from Transformers. 4, 8, 9, 52, 63, 64, 66, 73, 114, 116–121, 123, 126–128, 130, 135–137, 144, 155, 164, 165, 167, 168, 172, 175, XIII

Bi-LSTM Bidirectional Long Short Term Memory. 30, 118, 120

BIO Beginning, Inside, Outside. 27

CATCH Continuous Access To Cultural Heritage. 78

Corpus A large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based. 7, 8, 32, 36, 38, 40, 45, 60, 77, 81, 90, 115, 116, 120, 123, 124, 130, 131, 136, 159, 162, 163, 165, XIII

CRF Conditional Random Fields. 8, 30, 42, 43, 45, 76, 80, 82, 90, 114, 116–119, 123, 124, 126, 128, 135, 137, 164, 175

CSV Comma Separated Values. 145, 155

DANS Data Archiving and Networked Services. 2, 3, 19, 32, 38, 53, 54, 59, 77, 82, 83, 87, 90, 94–97, 115, 125, 140, 144, 153, 167, 169, 171, 173

Deep Learning A subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks. These Deep Learning algorithms are more complex and computationally expensive when compared to traditional Machine Learning approaches, but also generally produce better results. 32, 33, 165

DSRP Data Science Research Programme. 33, 77, 79

ElasticSearch An open source, full-text search engine. Used for all the indexing and retrieval tasks in this project. 76, 82, 83, 97, 116, 125, 126, 130, 145

EXALT EXcavating Archaeological LiTerature. 167, 170–172

F1 Score The F1 score (or F measure) combines recall and precision to provide an overall evaluation metric. More specifically, it is the harmonic mean of precision and recall. 6, 8, 28, 29, 34, 36, 37, 41–43, 45, 49, 51, 52, 63, 64, 66, 67, 69, 72, 74, 114, 116, 120, 121, 123, 124, 126, 128, 132, 136, 137, 145, 165, 166, 174

FAIR Findability, Accessibility, Interoperability, Reusability. 16, 20, 21, 158, 162, 163

GATE General Architecture for Text Engineering. 31, 79

GIS Geographical Information Systems. 90

Gold Standard A test set of human annotated documents describing the desirable system outcome. 31, 34, 82, 88–90

Grey Literature Materials and research produced by organisations outside of the traditional commercial or academic publishing and distribution channels. 2, 3, 8, 9, 16–18, 22, 29–32, 36, 48, 76–79, 87, 90, 92, 96, 110, 114, 140–143, 146, 154

HTTP Hypertext Transfer Protocol. 82

IAA Inter Annotator Agreement. 8, 36–39, 41, 121, 135, 165, 167, 174

Information Need A user's end goal in a specific search session, or a description of the information or the answer they are looking for. 8, 23, 24, 29, 48, 54, 90, 92, 93, 97, 100–104, 115, 116, 125, 137, 146, 148, 149, 154, 156

IR Information Retrieval. 22–24, 29–31, 76, 78, 86, 87, 92, 93, 95, 96, 114, 118, 119, 166, 174

JSON JavaScript Object Notation. 7, 82, 83, 125, 162

KB *Koninklijke Bibliotheek*. 19, 77, 90, 94, 167

LIACS Leiden Institute of Advanced Computer Science. 33
LOD Linked Open Data. 172

MEAN Miscellaneous, Exceptional, Arbitrary, Nonconformist. 162
Metadata Data that provides information about other data, often describing certain properties of a data set. 2, 6, 8, 21, 29, 30, 36, 48–50, 52–57, 59, 73, 77, 79, 80, 92, 94, 95, 97, 104, 114, 115, 119, 120, 125, 141, 153, 154, 161, 167, 173

NER Named Entity Recognition. 7, 8, 24–31, 33, 34, 36–38, 45, 50, 63, 73, 76, 78–82, 87–90, 97, 104, 114, 116–121, 123, 125, 126, 128, 130–132, 137, 141, 142, 144, 148, 153–155, 165, 166, 168, 171, 172, 174, 175
NLP Natural Language Processing. 22, 27, 29, 30, 51–53, 93, 118, 120, 144, 170
NOaA *Nationale Onderzoeksagenda Archeologie*. 162
NWO *Nederlandse organisatie voor Wetenschappelijk Onderzoek*. 19

OCR Optical Character Recognition. 15, 22, 38, 163

PDF Portable Document Format. 21, 31, 32, 36, 38, 49, 53, 83, 96, 97, 115, 141, 142, 163, 164, 173
POS Part Of Speech. 27, 81
Precision An evaluation measure that indicates, out of all the labelled entities, what percentage has been assigned the correct label. 6, 28, 29, 52, 64, 81, 82, 88, 92, 103, 115, 126, 128, 130, 134, 135, 148, 155, 156, 165, 166, 168
Python A widely used high-level programming language, used for most of the programming in this project. 81, 164

RCE *Rijksdienst voor het Cultureel Erfgoed*. 19, 33, 53, 77, 82, 85, 87, 89, 90, 94, 141, 161
RDF Resource Description Framework. 30
Recall An evaluation measure that indicates out of all the entities in a text, what percentage have been correctly labelled as an entity. 6, 24, 28, 29, 51, 52, 63, 64, 67, 69, 81, 82, 88, 92, 103, 115, 116, 125, 126, 128–130, 134, 137, 148, 165, 166, 168, 169, XII, XIII

SIKB *Stichting Infrastructuur Kwaliteitsborging Bodembeheer*. 19, 77, 173
SVM Support Vector Machine. 52, 62, 66, 67, 72, 164, 167

Text Mining The process of analysing text to extract information from it. 2, 3, 9, 12, 21, 24, 30, 31, 48, 76–79, 169
TF-IDF Term Frequency - Inverse Document Frequency. 126, 167

UI User Interface. 97, 109–111

XML eXtensible Markup Language. 34, 53–55, 81, 173