Knowledge extraction from archives of natural history collections

Stork, L.

Bibliography

[1] A. MacGregor, *Naturalists in the field: collecting, recording and preserving the natural world from the fifteenth to the twenty-first century*. Brill, 2018.

[2] P. L. Farber, *Finding order in nature: the naturalist tradition from Linnaeus to E.O. Wilson*. JHU Press, 2000.

[3] M. W. Holmes, T. T. Hammond, G. O. Wogan, R. E. Walsh, K. LaBarbera, E. A. Wommack, F. M. Martins, J. C. Crawford, K. L. Mack, L. M. Bloch, *et al.*, "Natural history collections as windows on evolutionary processes," *Molecular Ecology*, vol. 25, no. 4, pp. 864–881, 2016.

[4] M. Schilthuizen and F. Vonk, *Wie Wat Bewaart, die Heeft Wat*. Spectrum Uitgeverij Unieboek, 2020.

[5] T. Monquil-Broersen and E. Gassó, *Van onschatbare waarde: 200 jaar Naturalis*. Amsterdam University Press, 2021.

[6] A. H. Ariño, "Approaches to estimating the universe of natural history collections data," *Biodiversity Informatics*, vol. 7, no. 2, pp. 81–92, 2010.

[7] B. P. Hedrick, J. M. Heberling, E. K. Meineke, K. G. Turner, C. J. Grassa, D. S. Park, J. Kennedy, J. A. Clarke, J. A. Cook, D. C. Blackburn, S. V. Edwards, and C. C. Davis, "Digitization and the future of natural history collections," *BioScience*, vol. 70, no. 3, pp. 243–251, 2020.

[8] V. Blagoderov, I. J. Kitching, L. Livermore, T. J. Simonsen, and V. S. Smith, "No specimen left behind: industrial scale digitization of natural history collections," *ZooKeys*, vol. 209, pp. 133–146, July 2012.

[9] M. Heerlien, J. Van Leusen, S. Schnörr, S. de Jong-Kole, N. Raes, and K. Van Hulsen, "The natural history production line: an industrial approach to the digitization of scientific collections," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 8, no. 1, pp. 1–11, 2015.

[10] R. C. Baird, "Leveraging the fullest potential of scientific collections through digitisation.," *Biodiversity Informatics*, vol. 7, no. 2, 2010.

[11] E. Hyvönen, "Publishing and using cultural heritage linked data on the semantic web," *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 2, no. 1, pp. 1–159, 2012.

[12] V. Petras, T. Hill, J. Stiller, and M. Gäde, "Europeana–a search engine for digitised cultural heritage material," *Datenbank-Spektrum*, vol. 17, no. 1, pp. 41–46, 2017.

[13] N. E. Gwinn and C. Rinaldo, "The biodiversity heritage library: sharing biodiversity literature with the world," *IFLA journal*, vol. 35, pp. 25–34, March 2009.

[14] TDWG Interest Group, "Collection descriptions." `https://www.tdwg.org/community/cd/`, 2017. last accessed: 26-11-2020.

[15] Smithsonian Institution Archives, "The field book project." `https://siarchives.si.edu/about/field-book-project`, 2010. last accessed: 30-12-2020.

[16] A. Weber, "Collecting colonial nature: European naturalists and the netherlands indies in the early nineteenth century," *BMGN-Low Countries Historical Review*, vol. 134, no. 3, 2019.

[17] M. Moyle, J. Tonra, and V. Wallace, "Manuscript transcription by crowdsourcing: Transcribe bentham," *Liber Quarterly*, vol. 20, no. 3-4, 2011.

[18] K. A. Mika, J. De Veer, and C. Rinaldo, "Crowdsourcing natural history archives: Tools for extracting transcriptions and data," *Biodiversity Informatics*, vol. 12, pp. 58–75, 2017.

[19] A. Weber, M. Ameryan, K. Wolstencorft, L. Stork, M. Heerlien, and L. Schomaker, "Towards a digital infrastructure for illustrated handwritten archives," in *Digital Cultural Heritage* (M. Loannides, ed.), vol. 10605 of *Information Systems and Applications, incl. Internet/Web, and HCI*, pp. 155–166, Springer International Publishing, April 2018.

[20] R. E. Drinkwater, R. W. Cubey, and E. M. Haston, "The use of optical character recognition (ocr) in the digitisation of herbarium specimen labels," *PhytoKeys*, no. 38, p. 15, 2014.

[21] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger, "Transkribus-a service platform for transcription, recognition and retrieval of historical documents," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4, pp. 19–24, IEEE, 2017.

[22] J. A. Sánchez, V. Bosch, V. Romero, K. Depuydt, and J. De Does, "Handwritten text recognition for historical documents in the transcriptorium project," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pp. 111–117, 2014.

[23] L. Schomaker, "Design considerations for a large-scale image-based text search engine in historical manuscript collections," *It - Information Technology*, vol. 58, pp. 80–88, April 2016.

[24] T. M. Rath, R. Manmatha, and V. Lavrenko, "A search engine for historical manuscript images," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 369–376, 2004.

[25] H. S. Baird, V. Govindaraju, and D. P. Lopresti, "Document analysis systems for digital libraries: Challenges and opportunities," in *International Workshop on Document Analysis Systems*, pp. 1–16, Springer, 2004.

[26] M. Ameryan and L. Schomaker, "A high-performance word recognition system for the biological fieldnotes of the natuurkundige commissie.," in *Proceedings of the International Conference Collect and Connect (COLCO): Archives and Collections in a Digital Age*, pp. 92–103, 2020.

[27] L. Stork, A. Weber, J. van den Herik, A. Plaat, F. Verbeek, and K. Wolstencroft, "Large-scale zero-shot learning in the wild: Classifying zoological illustrations," *Ecological Informatics*, vol. 62, p. 101222, 2021.

[28] L. Stork, A. Weber, E. G. Miracle, F. Verbeek, A. Plaat, J. van den Herik, and K. Wolstencroft, "Semantic annotation of natural history collections," *Journal of Web Semantics*, vol. 59, 2019. 100462.

[29] J. B. Kennedy, R. Kukla, and T. Paterson, "Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration," in *International Workshop on Data Integration in the Life Sciences* (B. Ludäscher and L. Raschid, eds.), vol. 3615 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 80–95, Springer, 2005.

[30] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[31] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, pp. 49–79, December 2004.

[32] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[33] A. M. Lister and C. C. R. Group, "Natural history collections as sources of long-term datasets," *Trends in Ecology & Evolution*, vol. 26, pp. 153–154, January 2011.

[34] B. J. Cardinale, J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, G. M. Mace, D. Tilman, D. A. Wardle, *et al.*, "Biodiversity loss and its impact on humanity," *Nature*, vol. 486, no. 7401, pp. 59–67, 2012.

[35] A. Thomer, G. Vaidya, R. Guralnick, D. Bloom, and L. Russell, "From documents to datasets: A mediawiki-based metod of annotating and extracting species observations in century-old field notebooks," *ZooKeys*, vol. 209, pp. 235–253, July 2012.

[36] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais, "Darwin core: an evolving community-developed biodiversity data standard," *PloS one*, vol. 7, p. e29715, January 2012.

[37] E. G. Miracle, L. Stork, A. Weber, M. Ameryan, and K. Wolstencroft, "Natuurkundige commissie archives online." doi:10.1163/isbn.9789004336865, 2020. Leiden, the Netherlands: Brill.

[38] S. Müller-Wille, "Names and numbers:"data" in classical natural history, 1758–1859," *Osiris*, vol. 32, no. 1, pp. 109–128, 2017.

[39] E. G. Miracle, "On whose authority? temminck's debates on zoological classification and nomenclature: 1820–1850," *Journal of the History of Biology*, vol. 44, pp. 445–481, January 2011.

[40] W. G. Berendsohn, "The concept of "potential taxa" in databases," *Taxon*, vol. 44, pp. 207–212, May 1995.

[41] A. MacGregor, ed., *Naturalists in the Field*. Leiden, the Netherlands: Brill, 2018.

[42] G. E. Austen, M. Bindemann, R. A. Griffiths, and D. L. Roberts, "Species identification by experts and non-experts: comparing images from field guides," *Scientific Reports*, vol. 6, p. 33634, 2016.

[43] D. Maynard, K. Bontcheva, and I. Augenstein, "Natural language processing for the semantic web," *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 6, no. 2, pp. 1–194, 2016.

[44] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[45] R. C. Gonzales and R. E. Woods, "Digital image processing," 2002.

[46] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[48] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[49] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, pp. 4077–4087, 2017.

[50] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2013.

[51] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Advances in Neural Information Processing Systems*, no. 20, pp. 433–440, 2007.

[52] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[53] G. Antoniou, P. Groth, F. Van Harmelen, and R. Hoekstra, *A semantic web primer*. MIT press, 2004.

[54] D. K. Ahern, I. M. Braun, M. E. Cooley, and T. W. Bickmore, "Oncology informatics: behavioral and psychological sciences," in *Oncology informatics*, pp. 231–251, Elsevier, 2016.

[55] N. Guarino, D. Oberle, and S. Staab, "What is an ontology?," in *Handbook on ontologies*, pp. 1–17, Springer, 2009.

[56] T. R. Gruber, "Knowledge acquisition," *A translation approach to portable ontology specifications*, vol. 5, no. 199-220, pp. 10–1006, 1993.

[57] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.-C. Ngonga Ngomo, S. Rashid M., A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmerman, "Knowledge graphs," *arXiv preprint arXiv:2003.02320*, 2020.

[58] The Global Biodiversity Information Facility, "Gbif: The global biodiversity information facility (year) what is gbif?." `https://www.gbif.org/what-is-gbif`, 2020.

[59] GBIF Secretariat, "Gbif backbone taxonomy." `https://hosted-datasets.gbif.org/datasets/backbone/2018-06-20/`, 2018.

[60] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, IEEE, 2018.

[61] T. Causer and M. Terras, "'many hands make light work. many hands together make merry work': transcribe bentham and crowdsourcing manuscript collections," in *Crowdsourcing our cultural heritage*, pp. 57–88, Ashgate Farnham, 2014.

[62] V. De Boer, M. Hildebrand, L. Aroyo, P. De Leenheer, C. Dijkshoorn, B. Tesfa, and G. Schreiber, "Nichesourcing: harnessing the power of crowds of experts," in *International Conference on Knowledge Engineering and Knowledge Management*, pp. 16–20, Springer, 2012.

[63] C. Dijkshoorn, M. H. Leyssen, A. Nottamkandath, J. Oosterman, M. C. Traub, L. Aroyo, A. Bozzon, W. Fokkink, G.-J. Houben, H. Hovelmann, L. Jongma, J. van Ossenbruggen, G. Schreiber, and J. Wielemaker, "Personalized nichesourcing: Acquisition of qualitative annotations from niche communities.," in *UMAP Workshops*, 2013.

[64] M. Baechler, A. Fischer, N. Naji, R. Ingold, H. Bunke, and J. Savoy, "Hisdoc: historical document analysis, recognition, and retrieval," in *Digital humanities–international conference of the alliance of digital humanities organizations (ADHO)*, 2012.

[65] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character hmms," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 934–942, 2012.

[66] J. P. Philips and N. Tabrizi, "Historical document processing: Historical document processing: A survey of techniques, tools, and trends," *arXiv preprint arXiv:2002.06300*, 2020.

[67] L. Parilla and J. Blase, "The value of flexibility on long-term value of grant funded projects," *D-Lib Magazine*, vol. 21, no. 9/10, 2015.

[68] S. Nakasone and C. Sheffield, "Descriptive metadata for field books: Methods and practices of the field book project," *D-Lib Magazine*, vol. 19, p. 1, December 2013.

[69] T. Robertson, M. Döring, R. Guralnick, D. Bloom, J. Wieczorek, K. Braak, J. Otegui, L. Russell, and P. Desmet, "The gbif integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet," *PloS one*, vol. 9, August 2014. e102623.

[70] A. P. Giotis, G. Sfikas, B. Gatos, and C. Nikou, "A survey of document image word spotting techniques," *Pattern Recognition*, vol. 68, pp. 310–332, 2017.

[71] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, no. 2-4, pp. 139–152, 2007.

[72] N. Naji and J. Savoy, "Etude comparative de l'efficacité du dépistage de l'information dans des manuscrits médiévaux," in *Actes 11e Journées internationales d'analyse statistique des données textuelles JADT 2012*, pp. 753–766, 2012.

[73] A. H. Toselli, L. A. Leiva, I. Bordes-Cabrera, C. Hernández-Tornero, V. Bosch, and E. Vidal, "Transcribing a 17th-century botanical manuscript: Longitudinal evaluation of document layout detection and interactive transcription," *Digital Scholarship in the Humanities*, vol. 33, no. 1, pp. 173–202, 2018.

[74] A. H. Toselli, V. Romero, M. Pastor, and E. Vidal, "Multimodal interactive transcription of text images," *Pattern Recognition*, vol. 43, no. 5, pp. 1814–1825, 2010.

[75] V. Romero, A. H. Toselli, and E. Vidal, *Multimodal interactive handwritten text transcription*, vol. 80. World Scientific, 2012.

[76] S. Colutto, P. Kahle, H. Guenter, and G. Muehlberger, "Transkribus. a platform for automated text recognition and searching of historical documents," in *2019 15th International Conference on eScience (eScience)*, pp. 463–466, IEEE, 2019.

[77] A. Caceres, A. Weber, and L. Schomaker, "Monk in practice: Indexing heterogeneous handwritten collections," in *7th Digital Humatities Benelux 2020*, (Leiden, The Netherlands), 2020.

[78] E. Hyvönen, E. Heino, P. Leskinen, E. Ikkala, M. Koho, M. Tamper, J. Tuominen, and E. Mäkelä, "Warsampo data service and semantic portal for publishing linked open data about the second world war history," in *The Semantic Web. Latest Advances and New Domains* (H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, and C. Lange, eds.), (Cham), pp. 758–773, Springer International Publishing, 2016.

[79] V. de Boer, M. van Rossum, J. Leinenga, and R. Hoekstra, "Dutch ships and sailors linked data," in *International Semantic Web Conference (ISWC 2014)* (P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, eds.), vol. 8796 of *Lecture Notes in Computer Science*, (Cham), pp. 229–244, Springer International Publishing, October 2014.

[80] A. Meroño-Peñuela, A. Ashkpour, L. Rietveld, and R. Hoekstra, "Linked humanities data: The next frontier? a case-study in historical census data," in *Proceedings of the Second International Workshop on Linked Science 2012 - Tackling Big Data: in conjunction with the International Semantic Web Conference (ISWC2012)*, vol. 951, (Boston, MA), 2012.

[81] J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, and R. R. Swick, "Annotea: an open rdf infrastructure for shared web annotations," *Computer Networks*, vol. 39, no. 5, pp. 589–608, 2002.

[82] C. Dijkshoorn, V. De Boer, L. Aroyo, and G. Schreiber, "Accurator: Nichesourcing for cultural heritage," *Computing Research Repository*, 2017. abs/1709.09249.

[83] S. Ebert, M. Liwicki, and A. Dengel, "Ontology-based information extraction from handwritten documents," in *2010 12th International Conference on Frontiers in Handwriting Recognition*, pp. 483–488, IEEE, 2010.

[84] C. Adak, B. B. Chaudhuri, and M. Blumenstein, "Named entity recognition from unstructured handwritten document images," in *12th IAPR Workshop on Document Analysis Systems (DAS), 2016*, pp. 375–380, IEEE, 2016.

[85] J. I. Toledo, S. Sudholt, A. Fornés, J. Cucurull, G. A. Fink, and J. Lladós, "Handwritten word image categorization with convolutional neural networks and spatial pyramid pooling," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 543–552, Springer, 2016.

[86] M. Carbonell, M. Villegas, A. Fornés, and J. Lladós, "Joint recognition of handwritten text and named entities with a neural end-to-end model," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 399–404, IEEE, 2018.

[87] M. Greene and D. Meissner, "More product, less process: Revamping traditional archival processing," *The American Archivist*, vol. 68, no. 2, pp. 208–263, 2005.

[88] M. F. Desnoyers, "When is a collection processed?," *The Midwestern Archivist*, vol. 7, no. 1, pp. 5–23, 1982.

[89] TDWG Interest Group, "Minimum information about a digital specimen." `https://www.tdwg.org/community/cd/mids/`, 2017. last accessed: 26-11-2020.

[90] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM Sigmod Record*, vol. 34, no. 3, pp. 31–36, 2005.

[91] K. Eckert, "Provenance and annotations for linked data," in *International Conference on Dublin Core and Metadata Applications*, pp. 9–18, 2013.

[92] P. Groth, Y. Gil, J. Cheney, and S. Miles, "Requirements for provenance on the web," *International Journal of Digital Curation*, vol. 7, no. 1, pp. 39–56, 2012.

[93] D. Lewis, "General semantics," in *Montague grammar*, pp. 1–50, Elsevier, 1976.

[94] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, and H. van de Sompel, "The europeana data model (edm)," in *World Library and Information Congress: 76th IFLA general conference and assembly*, pp. 10–15, 2010.

[95] V. De Boer, J. Wielemaker, J. Van Gent, M. Hildebrand, A. Isaac, J. Van Ossenbruggen, and G. Schreiber, "Supporting linked data production for cultural heritage institutes: The amsterdam museum case study," in *The Semantic Web: Research and Applications. ESWC 2012.* (E. Simperl, P. Cimiano, A. Polleres, O. Corcho,

and V. Presutti, eds.), vol. 7295 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 733–747, Springer, 2012.

[96] C. Dijkshoorn, L. Aroyo, G. Schreiber, J. Wielemaker, and L. Jongma, "Using linked data to diversify search results a case study in cultural heritage," in *International Conference on Knowledge Engineering and Knowledge Management* (K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, eds.), vol. 8876 of *Lecture Notes in Computer Science*, (Cham), pp. 109–120, Springer International Publishing, 2014.

[97] M. Dragoni, E. Cabrio, S. Tonelli, and S. Villata, "Enriching a small artwork collection through semantic linking," in *The Semantic Web. Latest Advances and New Domains. ESWC 2016.* (H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, and C. Lange, eds.), vol. 9678 of *Lecture Notes in Computer Science*, (Cham), pp. 724–740, Springer International Publishing, 2016.

[98] M. Dragoni, S. Tonelli, and G. Moretti, "A knowledge management architecture for digital cultural heritage," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 10, pp. 1–18, August 2017.

[99] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, *et al.*, "Culturesampo: A national publication system of cultural heritage on the semantic web 2.0," in *European Semantic Web Conference*, pp. 851–856, Springer, 2009.

[100] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, "Methondology: from ontological art towards ontological engineering," in *Proceedings of the AAAI97 Spring Symposium*, pp. 33–40, March 1997.

[101] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López, "The neon methodology for ontology engineering," in *Ontology engineering in a networked world*, pp. 9–34, Springer, Berlin, Heidelberg, 2012.

[102] J. Gray and A. S. Szalay, "Where the rubber meets the sky: Bridging the gap between databases and science," *CoRR abs/cs/0502011*, 2005.

[103] S. J. Baskauf, J. Wieczorek, J. Deck, and C. O. Webb, "Lessons learned from adapting the darwin core vocabulary standard for use in rdf," *Semantic Web*, vol. 7, pp. 617–627, October 2016.

[104] S. J. Baskauf and C. O. Webb, "Darwin-sw: Darwin core-based terms for expressing biodiversity data as rdf," *Semantic Web*, vol. 7, pp. 629–643, October 2016.

[105] J. Tuominen, N. Laurenne, and E. Hyvönen, "Biological names and taxonomies on the semantic web–managing the change in scientific conception," in *The Semantic Web: Research and Applications. ESWC 2011.* (G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, eds.), vol. 6644 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 255–269, Springer, 2011.

[106] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel, "Uberon, an integrative multi-species anatomy ontology," *Genome biology*, vol. 13, p. R5, January 2012.

[107] G. Fragoso, S. de Coronado, M. Haber, F. Hartel, and L. Wright, "Overview and utilization of the nci thesaurus," *Comparative and Functional Genomics*, vol. 5, no. 8, pp. 648–654, 2004.

[108] M. Wick and B. Vatant, "The geonames geographical database." `http://www.geonames.org/`, 2012. last accessed: 30-03-2019.

[109] M. F. Loesch, "Viaf (the virtual international authority file)–http://viaf.org," *Technical Services Quarterly*, vol. 28, pp. 255–256, March 2011.

[110] B. Haslhofer, E. Momeni Roochi, B. Schandl, and S. Zander, "Europeana rdf store report," tech. rep., University of Vienna, 2011.

[111] E. Minack, W. Siberski, and W. Nejdl, "Benchmarking fulltext search performance of rdf stores," in *The Semantic Web: Research and Applications. ESWC 2009.* (L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, eds.), vol. 5554 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 81–95, Springer, 2009.

[112] R. Bhagdev, S. Chapman, F. Ciravegna, V. Lanfranchi, and D. Petrelli, "Hybrid search: Effectively combining keywords and semantic searches," in *The Semantic Web: Research and Applications* (S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, eds.), vol. 5021 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 554–568, Springer, 2008.

[113] E. Kaufmann, "Talking to the semantic web – query interfaces to ontologies for the casual user," in *The Semantic Web - ISWC 2006* (I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo, eds.), vol. 4273 of *Lecture Notes on Computer Science*, (Berlin, Heidelberg), pp. 980–981, Springer, November 2006.

[114] E. Kaufmann and A. Bernstein, "Evaluating the usability of natural language query languages and interfaces to semantic web knowledge bases," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 8, pp. 377–393, November 2010.

[115] D. A. Koutsomitropoulos, R. B. Domenech, and G. D. Solomou, "A structured semantic query interface for reasoning-based search and retrieval," in *The Semantic Web: Research and Applications. ESWC 2011.* (G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, eds.), vol. 6643 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 17–31, Springer, 2011.

[116] L. Stork and A. Kuzniar, "Sfb-annotator (version 0.1.1) zenodo." `https://doi.org/10.5281/zenodo.4602263`, 2021.

[117] A. Meroño-Peñuela and R. Hoekstra, "grlc Makes GitHub Taste Like Linked Data APIs," in *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 – June 2, 2016*, pp. 342–353, Springer, 2016.

[118] M. Ritsema van Eck and L. Schomaker, "Formal semantic modeling for human and machine-based decoding of medieval manuscripts," in *Proceedings of Digital Humanities*, pp. 336–338, University of Hamburg, July 2012.

[119] Z. Shi, "Datefinder: detecting date regions on handwritten document images based on positional expectancy," Master's thesis, University of Groningen, Groningen, the Netherlands, 2016.

[120] D. Koning, I. N. Sarkar, and T. Moritz, "Taxongrab: Extracting taxonomic names from text," *Biodiversity Informatics*, vol. 2, pp. 79–82, 2005.

[121] P. B. Heidorn and Q. Wei, "Automatic metadata extraction from museum specimen labels," in *International Conference on Dublin Core and Metadata Applications*, pp. 57–68, 2008.

[122] M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant, "Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 1, pp. 357–361, IEEE, 2007.

[123] T. Van der Zant, L. Schomaker, and K. Haak, "Handwritten-word spotting using biologically inspired features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1945–1957, 2008.

[124] J.-P. van Oosten and L. Schomaker, "Separability versus prototypicality in handwritten word-image retrieval," *Pattern Recognition*, vol. 47, no. 3, pp. 1031–1038, 2014.

[125] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[126] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[127] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[128] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, 2009.

[129] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724, IEEE, 2014.

[130] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 188–191, Association for Computational Linguistics, 2003.

[131] J. A. Drew, C. S. Moreau, and M. L. Stiassny, "Digitization of museum collections holds the potential to enhance researcher diversity," *Nature ecology & evolution*, vol. 1, no. 12, p. 1789, 2017.

[132] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[133] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, IEEE, 2015.

[134] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2011–2018, IEEE, 2014.

[135] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 951–958, IEEE, 2009.

[136] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. B. Soares, "Leafsnap: A computer vision system for automatic plant species identification," in *Proceedings of the European Conference on Computer Vision*, pp. 502–516, Springer, 2012.

[137] M. E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1447–1454, IEEE, 2006.

[138] S. Beery, E. Cole, and A. Gjoka, "The iwildcam 2020 competition dataset," *arXiv preprint arXiv:2004.10340*, 2020.

[139] O. Mac Aodha, E. Cole, and P. Perona, "Presence-only geographical priors for fine-grained image classification," in *Proceedings of the IEEE International Conference on Computer Vision*, (Seoul, Korea (South)), pp. 9595–9605, 2019.

[140] G. Chu, B. Potetz, W. Wang, A. Howard, Y. Song, F. Brucher, T. Leung, and H. Adam, "Geo-aware networks for fine-grained recognition," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, (Seoul, Korea (South)), pp. 247–254, 2019.

[141] S. Beery, G. Wu, V. Rathod, R. Votel, and J. Huang, "Context r-cnn: Long term temporal context for per-camera object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Seattle, Washington), pp. 13072–13082, 2020.

[142] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Fine-grained object recognition and zero-shot learning in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 770–779, 2018.

[143] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, IEEE, 2015.

[144] P. N. Belhumeur, D. Chen, S. Feiner, D. W. Jacobs, W. J. Kress, H. Ling, I. Lopez, R. Ramamoorthi, S. Sheorey, S. White, and L. Zhang, "Searching the world's herbaria: A system for visual identification of plant species," in *Proceedings of the European Conference on Computer Vision*, pp. 116–129, Springer, 2008.

[145] B. Barz and J. Denzler, "Hierarchy-based image embeddings for semantic image retrieval," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 638–647, IEEE, 2019.

[146] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, no. Sep, pp. 1453–1484, 2005.

[147] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, vol. 2, pp. 3111–3119, 2013.

[148] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, Association for Computational Linguistics, 2014.

[149] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[150] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015.

[151] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems*, pp. 2121–2129, 2013.

[152] B. Romera-Paredes and P. H. S. Torr, "An embarrassingly simple approach to zero-shot learning," in *Visual Attributes*, pp. 11–30, Springer, 2017.

[153] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 69–77, 2016.

[154] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in Neural Information Processing Systems*, pp. 935–943, 2013.

[155] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758, IEEE, 2012.

[156] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[157] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning–a comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2019.

[158] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, IEEE, 2016.

[159] N. T. H. Nguyen, A. J. Soto, G. Kontonatsios, R. Batista-Navarro, and S. Ananiadou, "Constructing a biodiversity terminological inventory," *PLoS ONE*, vol. 12, no. 4, p. e0175277, 2017.

[160] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Workshop proceedings of the International Conference on Learning Representations*, arXiv preprint arXiv:1301.3781, 2013.

[161] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of Artificial Intelligence Research*, vol. 37, no. 1, pp. 141–188, 2010.

[162] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[163] P. M. Choate, "Introduction to the identification of beetles (coleoptera)," *Dichotomous keys to some Families of Florida Coleoptera*, pp. 23–33, 1999.

[164] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[165] M. Ameryan and L. Schomaker, "Improving the robustness of lstms for word classification using stressed word endings in dual-state word-beam search," in *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 13–18, IEEE, 2020.

[166] S. Chanda, J. Baas, D. Haitink, S. Hamel, D. Stutzmann, and L. Schomaker, "Zero-shot learning based approach for medieval word recognition using deep-learned

features," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 345–350, IEEE, 2018.

Acronyms

AI artificial intelligence. 15, 118, 122, 153, 157

ALICE Artificial Intelligence and Cognitive Engineering. iv, 13

ANN artificial neural network. 25, 26, 27, 81

API Application Programming Interface. 35, 63, 64, 74

BHL Biodiversity Heritage Library. 35, 61, 78, 79, 98, 101, 105, 110, 117, 118, 121

BLSTM bilateral long short-term memory network. 28, 43, 45, 82, 83, 84, 86, 87

CE Concatenated Embedding. 101

CH cultural heritage. 50

CNN convolutional neural networks. 27, 28, 81, 82, 84, 85, 86, 87, 94, 99

COL Catalogue of Life. 117

DC Dublin Core. 61, 62

DIAR Document Image Analysis and Recognition. 77, 79

DNN deep (artificial) neural network. 25, 26, 28

DSW Darwin Core Semantic Web. 53, 54, 56, 57, 58, 83

DwC Darwin Core. 31, 41, 42, 53, 54, 56, 57, 58, 61, 62, 63, 64, 79

EOL Encyclopedia of Life. 35

FAIR Findable, Accessible, Interoperable, and Reusable. 8, 10, 31, 48, 49, 120, 121, 122, 148, 150

FOAF Friend Of A Friend. 59

FP Fused Prototype. 92, 98, 102, 116

FSL few-shot learning. 28, 29, 102

GBIF Global Biodiversity Data Facility. 35, 42, 62, 96, 98, 101, 105, 117, 121

GLAMs Galleries, Libraries, Archives and Museums. 37, 40, 42, 45

GZSL generalised zero-shot learning. 94, 95, 112, 114

HMM Hidden Markov Model. 43

HPL Hierarchical Prototype Loss. 92, 98, 102, 103, 112, 116

HTR Handwritten Text Recognition. 6, 13, 38, 39, 41, 43, 44, 45, 46, 77, 78, 79, 81, 88, 121, 122

IIIF International Image Interoperability Framework. 72, 74

IRI Internationalised Resource Identifier. 8, 33, 39, 48, 50, 52, 53, 54, 60, 62, 63, 65, 67, 68, 85

IUCN International Union for Conservation of Nature and Natural Resources. 35

JSON JavaScript Object Notation. 64

KRR knowledge representation and reasoning. 22, 30, 119

LCDS Leiden Centre of Data Science. iv

LD Linked Data. 30, 48

LIACS Leiden Institute of Advanced Computer Science. iv, 13, 153, 157, 158

LINNAE LInking Notes of NAturE. 11, 72, 159

LOD Linked Open Data. 11, 120

LSTM long short-term memory network. 28, 83

MIDS Minimum Information about a Digital Specimen. 47

MLP multi-layer perceptron. 26, 27, 28, 82, 84, 86, 87

MODS Metadata Object Description Schema. 61

NBC Naturalis Biodiversity Center. iv, 1, 2, 3, 13, 18, 36, 54

NC Committee for Natural History of the Netherlands Indies ("Natuurkundige Commissie voor Nederlands-Indië"). 1, 2, 13, 20, 35, 51, 66, 72, 88

NCD Natural Collections Description. 41, 61, 62

NCO Natural Committee Online. 13

NER named entity recognition. 45

NERC named entity recognition and classification. 7, 11, 22, 24, 44, 46, 48, 75, 78, 79, 82, 120, 121

NHC natural history collection. i, 3, 5, 6, 7, 9, 10, 11, 12, 13, 17, 34, 47, 49, 54, 56, 57, 60, 61, 62, 63, 68, 69, 77, 78, 81, 83, 87, 89, 117, 118, 119, 120, 121, 122, 147

NLP natural language processing. 38, 44, 75

NMNH Smithsonian National Museum of Natural History. 61

OBIE ontology-based information extraction. 45

OCR Optical Character Recognition. 6, 35, 39, 43, 79

OOV out-of-vocabulary. 6

OWL Web Ontology Language. 33, 50, 52, 83

PNL Prototypical Network Loss. 99, 103, 112

RDF Resource Description Framework. 8, 33, 34, 45, 53, 59, 64, 74, 85, 87, 88

RNN recurrent neural network. 28

ROI region of interest. 39, 40, 43, 46, 63, 64, 67

SFB-Annotator Semantic Field Book Annotator. 11, 50, 61, 63, 65, 66, 71, 72, 74, 117, 120, 122, 148, 150

SGD Stochastic Gradient Descent. 99

SIA Smithsonian Institution Archives. 61

SNERC salient named entity recognition and classification. 11, 12, 52, 65, 78, 120, 122

SPARQL SPARQL Protocol and RDF Query Language. 8, 34, 68, 69, 71, 72, 74, 87

STePS Department of Science, Technology, and Policy Studies. iv, 13

SVM Support Vector Machine. 107, 108

t-SNE t-Distributed Stochastic Neighbour Embedding. 106, 107, 115

TDWG Biodiversity Information Standards. 35, 47

TEI Text Encoding Initiative. 41

TNS Taxonomic Name Server. 79

URI Uniform Resource Identifier. 8, 21, 33, 34, 46

URL Uniform Resource Locator. 8, 33

VGG Visual Geometry Group. 81, 82

VIAF Virtual International Authority File. 48, 59, 65

W3C World Wide Web Consortium. 33, 52, 60

XML Extensible Markup Language. 31, 32, 41, 42, 61

ZICE Zoological Illustration and Class Embedding. 12, 92, 95, 97, 98, 106, 115

ZSL zero-shot learning. 6, 12, 25, 28, 29, 91, 92, 93, 94, 95, 96, 98, 99, 105, 109, 111, 112, 114, 116, 121, 122, 151