Biographical Data in a Digital World 2022 (BD 2022) Workshop

Daza, Angel

Daza , A , Fokkens , A , Hadden , R , Hyvönen , E , Koho , M & Wandl-Vogt , E 2022 ,
Biographical Data in a Digital World 2022 (BD 2022) Workshop . in Digital Humanities 2022 :
Conference Abstracts: The University of Tokyo, Japan 25-29 July 2022 . The Alliance of
Digital Humanities Organizations (ADHO) , pp. 39-42 , Digital Humanities , Tokyo , Japan ,
25/07/2022 . < https://dh2022.adho.org/ >

unspecified
publishedVersion

Instructors

- **Mariona Coll Ardanuy:** Mariona is a computational linguist at the Alan Turing Institute in the Living with Machines project. Her research interests lay in the intersection between the humanities and language technology.
- **Kasra Hosseini:** Kasra is a Research Data Scientist at The Alan Turing Institute. He is interested in (artificially) intelligent systems, machine learning, and data analysis and visualisation.
- **Federico Nanni:** Federico is a Research Data Scientist at The Alan Turing Institute. He is a historian by training and works exploring the intersections between digital humanities, computational social science, and natural language processing.
- **Valeria Vitale:** Valeria Vitale is a researcher in the field of digital cultural heritage. She works at the Alan Turing Institute as Research Associate on the Machines Reading Maps project.

Target audience

Based on past experience, we believe the number of participants should be 20 at most. Participants should have some experience in programming in Python and running scripts, and ideally be interested in entity linking or fuzzy string matching.

Funding statement

Bibliography

Butler, J. O., Donaldson, C. E., Taylor, J. E., and Gregory, I. N. (2017). Alts, Abbreviations, and AKAs: historical onomastic variation and automated named entity recognition. *Journal of Map & Geography Libraries*, *13*(1), 58-81.

Coll Ardanuy, M., Hosseini, K., McDonough, K., Krause, A., van Strien, D., and Nanni, F. (2020). A deep learning approach to geographical candidate selection through toponym matching. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems* (pp. 385-388).

Coll Ardanuy, M., and Sporleder, C. (2017). Toponym disambiguation in historical documents using semantic and geographic features. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage* (pp. 175-180).

De Wilde, M., and Hengchen, S. (2017). Semantic enrichment of a multilingual archive with linked open data. *Digital Humanities Quarterly*.

Evershed, J., & Fitch, K. (2014). Correcting noisy OCR: Context beats confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (pp. 45-51).

Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with Wikipedia. *Artificial intelligence*, *194*, 130-150.

Hosseini, K., Nanni, F., and Coll Ardanuy, M. (2020). DeezyMatch: A flexible deep learning approach to fuzzy string matching. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 62-69).

Olieman, A., Beelen, K., van Lange, M., Kamps, J., and Marx, M. (2017). Good applications for crummy entity linkers? the case of corpus selection in digital humanities. In *Proceedings of the 13th International Conference on Semantic Systems* (pp. 81-88).

Santos, R., Murrieta-Flores, P., and Martins, B. (2018). Learning to combine multiple string similarity metrics for effective toponym matching. *International journal of digital earth*, *11*(9), 913-938.

Santos, R., Murrieta-Flores, P., Calado, P., and Martins, B. (2018). Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, *32*(2), 324-348.

Van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the impact of OCR quality on downstream NLP tasks. *Special Session on Artificial Intelligence and Digital Heritage: Challenges and Opportunities*, in *Proceedings of the 12th International Conference on Agents and Artificial Intelligence* (pp. 484-496)

Biographical Data in a Digital World 2022 (BD 2022) Workshop

Daza, Angel

j.a.dazaarevalo@vu.nl
Vrije Universiteit Amsterdam, The Nethelands

Fokkens, Antske

antske.fokkens@vu.nl
Vrije Universiteit Amsterdam, The Nethelands; Eindhoven University of Technology, The Nethelands

Hadden, Richard

richard.hadden@oeaw.ac.at
Austrian Academy of Sciences, Austria

Hyvönen, Eero

eero.hyvonen@aalto.fi
Aalto University, Finland; Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland

Koho, Mikko

mikko.koho@aalto.fi
Aalto University, Finland

Wandl-Vogt, Eveline

eveline.wandl-vogt@oeaw.ac.at
Ars Electronica Research Institute, Austria; Austrian Academy of Sciences, Austria

Biographical and prosopographical (Verboven et al., 2007) data are invaluable sources for historical research as they provide us with essential information on thousands of historical figures: from the cultural heroes of a nation to the many thousands of other significant, yet lesser-known figures who were influential in domains such as the arts, politics, humanities, or natural sciences. Their historical life paths can provide crucial context for tangible heritage objects, which have been created, owned, or influenced by historical actors, or which depict or refer to them. The events of individual biographies can further be aggregated into (histories of) larger contextual composites: groups (e.g., guilds, family histories), institutions (e.g., art schools, universities, religious orders, political movements, companies) and regional entities (from cities to whole countries).

Computational analysis of biographies has opened up new and interesting research directions (Warren, 2018; Fokkens et al., 2017; Tamper et al., 2022). Individuals share common characteristics that can be relatively easily identified and used for information retrieval and data analysis, such as date of birth, occupation, social networks, and locations. Tools and approaches from the digital humanities can be used for both quantitative analyses of

such data and for providing leads for more qualitative research questions. New methods of spatial analysis may result in new research questions.

Workshop Goals

This workshop has two main goals. First, previous events on biographical data and digital humanities have revealed that groups working in this domain have different strengths. Though many groups bring researchers from different domains together, some groups mainly consist of domain experts with a lot of knowledge of history, library studies or literature. Other teams are primarily made up of researchers specialized in automatic analysis, formal modeling or visualization. Bringing these groups together increases insights for both types of groups. Technical experts contribute insights on best practices to deal with typical challenges of the domain. Domain experts on what analyses they need, which potential technological challenges are problematic for their research questions and which are not. Second, there is keen interest in cross-national research on biographical information. Most resources are national resources and though many are open, sharing information also means understanding each other's data representation, linking related information, etc.

The *Biographical Data Workshop: modeling, sharing and analyzing people's lives* held at Digital Humanities 2016 in Krakow led to valuable steps on both these points. Among others, the first steps were made that ultimately led to active international collaboration (e.g., the InTaVia project: https://intavia.eu/). Our motivation for organizing this follow-up is two-fold. First, the advances of methods and availability of new resources in the last six years leads us to expect that new interactive sessions of knowledge exchange will have the same beneficiary effect as previous events. Second, the workshop and *Biographical Data in a Digital World* conference series were, in terms of participants, rather European centric with some participants from America. It is our hope that a workshop at the current location will strengthen connections with researchers from Asia. In addition to the general increase of the network of expertise in this domain, steps towards connecting resources for Asia would open new horizons for new comparative research.

Workshop Organization

The workshop will be divided into three sessions.

Preliminary schedule

1. (morning) The participating researchers get the chance to present their work. Depending on the number of submissions, this will be done through (short) presentations or in two poster sessions (where we aim to schedule posters of researchers of existing networks in the same session, to increase options for new connections).
2. (afternoon) The session consists of two interactive themed sessions. In the first round, we have technically oriented themes (data linking and sharing, data analysis and data visualization), where technical experts meet with various domain experts. In the second round, sessions are organized around use cases from the humanities, where humanity scholars meet with a team with different technical expertise.
3. (afternoon) Work groups will share their insights from the interactive sessions with the rest of the participants.

Submissions, proceedings, and workshop participation

We solicit abstract submissions for the poster session through its own call for participation. Possibilities of peer-reviewed proceedings based on the workshop will be discussed with participants during the workshop.

The contributions will be submitted through the main conference ConfTool system: https://www.conftool.pro/dh2022/.

At least one author of each paper needs to register to the DH 2022 conference in order to participate in the workshop. The workshop is open for all conference participants. We reach out for close collaboration with the ADHO Geohumanities SIG (https://geohumanities.org/).

Logistics

Important dates

- March 4: Call for abstracts and requests for participation
- April 14: Deadline for abstracts
- April 29: Notification of acceptance (peer-review results)
- July 25/26: The workshop takes place

Intended audience

The target audience of the workshop are researchers working with biographical or prosopographical data. The audience is expected to consist of DH-focused researchers and computer scientists, as well as social science and humanities researchers.

Expected number of participants

Estimating the number of participants physically present is difficult due to COVID, but the previous BD workshops have had dozens of attending participants; there will be more with the planned possibility for online access.

Intended length

One day.

Budget

The workshop is self-financing.

Topics

Topics of interest include, but are not limited to:

- Digitizing and structuring biographical data
- Standards, vocabularies and best practices for processing biographical data
- Biographies and Linked Data
- Crowdsourcing biographical data
- Automatic biography generation
- Using biographical and prosopographical data for quantitative analyses
- Canonization of people and events in history
- Use of big data for biographical research
- Dealing with biographical data in heterogeneous datasets
- Creating and maintaining biographical dictionaries
- Enriching biographies from external sources
- Reconciling persons between biographical dictionaries
- Reconciling names against a biographical dictionary
- Visualizing biographical and prosopographical data
- Network analysis of biographical data
- Biographies and spatial analysis
- Biographies across countries and cultures

Organizers

Angel Daza, Vrije Universiteit Amsterdam (j.a.dazaarevalo@vu.nl). Angel Daza is a computer scientist with expertise in Natural Language Processing. He completed his Ph.D. research at Universität Heidelberg

where he worked on multilingual models for Semantic Role Labeling. He recently joined as a postdoc for the Computational Linguistics & Text Mining Lab at VU and is part of the In/Tangible European Heritage Visual Analysis, Curation & Communication project (InTaVia).

Antske Fokkens, VU University Amsterdam and Eindhoven University of Technology (antske.fokkens@vu.nl) holds a University Research Chair on Computational Linguistic Methods. The chair specifically focuses on methodological aspects of language technologies when they are used in other disciplines (such as history, literature and library studies). She was one of the main researchers in the BiographyNet project and currently leads the Text Mining package of the InTavia project. Together with Historian Serge ter Braake she initiated the Biographical Data in a Digital World conference series and is co-chair of the DARIAH EU working group analyzing and linking biographical data.

Richard Hadden, Austrian Academy of Sciences (richard.hadden@oeaw.ac.at). Richard Hadden is a Digital Humanities researcher. He completed a PhD in textual scholarship and large-scale digitisation and knowledge representation at Maynooth University, Ireland. He joined the Austrian Centre for Digital Humanities and Cultural Heritage in 2020, as a postdoctoral researcher in digital prosopography. He currently works on several prosopographical and biographical projects as the ACDH-CH, including development of the IPIF prosopographical data format.
https://www.oeaw.ac.at/acdh/team/current-team/richard-hadden

Eero Hyvönen, Helsinki Centre for Digital Humanities (HELDIG) and Aalto University (eero.hyvonen@aalto.fi). Eero Hyvönen is professor of semantic media technology in Aalto University and director of the Helsinki Centre for Digital Humanities (HELDIG); he has been involved in developing several biographical and prosopographical systems for DH research, based on Semantic Web technologies, Linked Data, and Artificial Intelligence.
https://seco.cs.aalto.fi/u/eahyvone/

Mikko Koho, Aalto University (mikko.koho@aalto.fi). Mikko Koho is a Staff Scientist at Aalto University, Department of Computer Science. Research focuses on Linked Data, ontologies, and data modelling, as well as data analysis in multidisciplinary Digital Humanities research.
https://seco.cs.aalto.fi/u/mkoho/

Eveline Wandl-Vogt, (eveline.wandl-vogt@oeaw.ac.at) foundress and director of Ars Electronica Research Institute knowledge for humanity, research manager and experimental researcher at Austrian Academy of Sciences, affiliate at metalab at Harvard, foundress and co-chair of DARIAH EU working group analyzing and linking biographical data, network and open innovation facilitator eg for biographical and prosopographical data,

is challenging the genre against an art driven and open innovation background.
https://www.linkedin.com/in/evelinewandlvogt

Confirmed Program Committee members

- Paul Arthur, Edith Cowan University, Australia
- Angel Daza, Vrije Universiteit Amsterdam, The Netherlands
- Thierry Declerck, German Research Center for Artificial Intelligence (DFKI), Germany
- Antske Fokkens, Vrije Universiteit Amsterdam, The Netherlands
- Richard Hadden, Austrian Academy of Sciences, Austria
- Eero Hyvönen, University of Helsinki and Aalto University, Finland
- Mikko Koho, Aalto University, Finland
- Lik Hang Tsui, City University of Hong Kong, Hong Kong
- Eveline Wandl-Vogt, Ars Electronica Research Institute and Austrian Academy of Sciences, Austria
- Hongsu Wang, Harvard University, USA

Bibliography

Fokkens, A. S., ter Braake, S., Ockeloen, C. J., Vossen, P. T. J. M., Legêne, S., Schreiber, G. and de Boer, V. (2017). BiographyNet: Extracting Relations between People and Events. In Bernád, Á. Z., Gruber, C. and Kaiser, M. (eds), Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik. Wien: New Academic Press, pp. 193-224.

Tamper, M., Leskinen, P., Hyvönen, E., Valjus, R. and Keravuori, K. (2022): Analyzing Biography Collection Historiographically as Linked Data: Case National Biography of Finland. Semantic Web. Accepted for publication.

Verboven, K., Carlier, M. and Dumolyn, J. (2007). A Short Manual to the Art of Prosopography. In Keats-Rohan, K.S.B. (ed), Prosopography Approaches and Applications. A Handbook. Oxford: Occasional Publications of the Unit for Prosopographical Research, pp. 35-70.

Warren, C. N. (2018). Historiography's Two Voices: Data Infrastructure and History at Scale in the Oxford Dictionary of National Biography (ODNB). Journal of Cultural Analytics, 3(1).
https://doi.org/10.22148/16.028 (accessed 28 April 2022).