CLARIAH Research Pilot Final report

The research pilot final report is written by the project coordinator and approved by the project participants. Next to this report, a separate financial report has to be made and signed by each participating organization, and the project coordinator makes an additional financial report for the project as a whole. Submit all reports via e-mail to clariah@huygens.knaw.nl.

General

Project name: The History of Human Capital

Project acronym: HHuCap

Project code: 300-2-0021
Project website: https://www.clariah.nl/en/projects/research-pilots/granted-pilot-research-projects/hhucap/hhucap

Project coordinator: Richard Zijdeman

Project budget: €79.660,-

Project summary: include the summary of the project proposal

Target Start Date	Actual Start Date	Target End Date	Actual End Date
1-1-2018	12-1-2018	31-12-2018	03-02-2020

Contact person

Name	Richard Zijdeman
Organisation	International Institute of Social History
Street	Cruquiusweg
Number	31
Postal code	1019 AT
City	Amsterdam
Telephone	+31 20 668 5 866
E-mail	richard.zijdeman@iisg.nl

Project partners

Organisation (acronym)	Organisation name	City	From (date)	To (date)
VU	Vrije Universiteit Amsterdam	Amsterdam	1-1-2018	1-3-2020
UU	Universiteit Utrecht	Utrecht	1-1-2018	1-3-2020
IISG	IISG	Amsterdam	1-1-2018	1-3-2020

Deliverables

ID	Type	Title	Target Date	Revised Target Date	Delivery Date	Remarks[1]
1	Dataset	BP Evaluation	05-2018		04-2018	https://github.com/CLARIAH/hhucap/tree/master/figures
2	Dataset	KB Job Advertisements	09-2018		07-2018	Available on request (user rights required by KB) via auke.rijpma@uu.nl
6	Dataset	Biography Annoted Data	10-2018		07-2018	https://github.com/CLARIAH/hhucap/tree/master/data/bio-xml-output
11	Doc	Final Report	12-2018		12-2019	[This document]
12	Doc	eHost Manual and FAQ	-		12-2018	https://github.com/CLARIAH/hhucap/blob/master/docs/handleiding_eHost.pdf
13	Doc	Guidelines occupation annotation	-		05-2018	https://github.com/CLARIAH/hhucap/blob/master/docs/guidelines_annotation_occupation_nl.md
14	Dataset/API	HISCO	-		09-2018	https://druid.datalegend.net/dataLegend/HISCO/

Remarks

Deliverables 3 to 5 and 7 to 10 have not been produced. In short, the evaluation of the two components (manual and automated tagging) proved so complicated to setup, that the project did not allow for the evaluation. Instead, we provided two new deliverables 12 to 14. For more detailed comments see the section “Deviations from the Original Plan”.

Deviations from the Original Plan

From a NLP perspective, the idea behind the project was to evaluate a machine-annotated and human-annotated corpus for occurrences of occupations and their

[1] **ID**= identifier of the Deliverable, **Type** = document, software, data, milestone

incumbents from the perspective of the bio-subject. From a research perspective the idea behind the project was to use the best set of annotations to describe careers. This would be novel as biographies provide much more detailed career descriptions than more commonly used sources, such as linked event registers.

For the annotation of our corpus by humans we thought we were able to rely on an CLARIAH pipeline (as this is also the tool we wanted to evaluate (deliverables 3,8)). However, such a pipeline did not exist, nor a CLARIAH supported annotation tool. To still execute deliverable 6 we relied instead on eHOST. We lost quite some time due to this detour, not just because of the alternate strategy, but also because eHOST had to be installed on students machines (most of them running windows, requiring them to set up a VM). eHOST came without proper user documentation, which is what we provided in unforeseen deliverable 12.

What we grossly underestimated is the element of creating guidelines for human annotation of occupations and their incumbents (deliverable 13). We failed to explicitly plan such a full-fledged version, and it also took time to adjust the wordings in such a way, that all human annotators interpreted them in the same way. As documentation and for the purpose of preparing others for this job, we have added the nearly 40 questions as FAQ to the eHOST documentation.

To allow for subsequent projects to more easily code Dutch occupations into HISCO, the Historical Standard Classification of Occupations, we have provided a Linked Data version of HISCO, which can be reused in future HISCO-tagger pipelines.

Another aspect that didn't quite help was that the staff that was supposed to work on the project was pulled out, amongst others for unforeseen teaching tasks. As a result, we had to form a team of student assistants. What worked quite well, is that we tasked two-master students to co-supervise four bachelor students. This did mean, that non-project time was devoted to supervise all students and to steer the project in its alternate direction. We thank the IISG, UU and VU for providing this additional time.

The HHuCap project had positive outcomes that extend beyond the project. Amongst others because of his work on the HHuCap project, Ruben Ros, one of the master student-assistants, attained an intership on the derivation of wages from newspaper articles with Marieke van Erp at HuC-DHLab. Also, through the HHuCap project a collaboration with CLARIAN.eu project LR4SSHOC was established. The results from the process of text-mining Social Economic History data from newspapers will be presented at the LREC2020 workshop, May 11, and published as workshop proceedings.

Dissemination

Send a PDF version of each publication, presentation, report, etc. to clariah@huygens.knaw.nl for inclusion on the CLARIAH website and for project reporting purposes.

Provide the proper bibliographic references for each form of dissemination below.

Website, Wiki, etc.

https://github.com/CLARIAH/hhucap

Presentations

Zijdeman, R.L. 2019. Linked Data: Een extra ontsluitingslaag op Archieven. Digital Humanities & Archieven Workshop. March 5th. Nationaal Archief, Den Haag.

Zijdeman, R.L. 2018. HHuCap Elevator Pitch. CLARIAH Toogdag. March 9th. Koninklijke Bibliotheek, Den Haag.

Fokkens, A., Zijdeman, R.L. & A. Rijpma. 2017. Distilling Careers: augmenting biographies with occupational information. Digital Humanities Benelux, July 5-9, Utrecht University, Utrecht. https://github.com/CLARIAH/hhucap/blob/master/presentations/hhucap_demo_dhb2017.pdf

Internal Reports

Ros, R. & D. Braven. 2018. Handleiding Annotation tool eHost. https://github.com/CLARIAH/hhucap/docs/

Ros, R. & D. Braven. 2018. Guidelines for annotating occupations. https://github.com/CLARIAH/hhucap/docs/

Tutorials or other educational activities organised

Fokkens, A. & I. Maks. 2018. Workshop eHost Annotation tool. February 22nd, VU University. Session hosted for the VU/UU students in the project.

Ros, Ruben. 2019. Mining wages in nineteenth century newspaper job advertisements. Internship at HuC-DHLab. Supervision by Van Erp, M., Rijpma. A & R.L. Zijdeman

Other Activities

Results of the process of the automated textmining procedure will be presented and appears as proceedings (tbc) at: Ros, R., Van Erp, M., Rijpma, A. & R.L. Zijdeman. 2020. LREC2020 workshop: Language for the Cloud. Marseille, May 11, 2020.

Recommendations for CLARIAH

The opportunity for cross-workpackage cooperation through the HHuCap project has had important benefits. For one, the project has created a better understanding of the technical and methodological difficulties that come with computational analysis of text in the domain of social and economic history. In addition, while the proposed evaluation paper of manual and machine annotation proved beyond the scope of the project, there is a clear sense of the difficulties in both approaches. An important result is the realization that human inter-annotation scores are not always driven by substantial differences in annotation, and not less error prone than machine

annotated materials. The recommendation therefore is to continue providing ‘seed money’ especially between cross-workpackage projects.

A major difficulty we encountered was that there was no CLARIAH annotation tool that we could use from the start. There were several projects under way, but none were easily installable to laymen (or could be used as a service). In future calls, we would therefore recommend to highlight the tools that are ready and available for evaluation along with a devoted person, that can help overcome technical difficulties with the CLARIAH tools.