1. Project Title & Acronym and Abstract

Title	The History of Human Capital
Acronym	HHu'Cap
Abstract	Current historical studies of career mobility often focus on linkage of personal records such as baptism records. More qualitative sources, such as biographies contain vital information as well, but are labour intensive to process. We propose a combination of Robust Semantic Parsing and Linked Data conversion tools to automatically derive career patterns from 35,000 biographies in the Biography Portal in the period 1815-1940. Substantively, we answer the question what career patterns looked like and changed over the long Nineteenth century. Methodologically, we evaluate to what extent current CLARIAH tools are up to automate this process. We will progress the semantic parsing tools by improving the linguistic expression set related to HISCO, adding an OCR cleaning step to the pipeline and experimenting with alternative CLARIAH tools for Dutch. This will result in a detailed report on the performance of CLARIAH tools on this data.
Target Start Date	1 May 2017
Target End Date	30 April 2018

2. Principal Investigator

Name	Richard Zijdeman
Function	Senior Researcher, Chief Data Officer
Organization	International Institute of Social History
Address	IISH, Cruquiusweg 31, 1019 AT, Amsterdam
E-mail	richard.zijdeman@iisg.nl
Tel	+31 20 66 858 66

3. Composition of the Project Team

Name	Antske Fokkens
Function	Researcher
Organization	Vrije Universiteit
Address	De Boelelaan 1105 1081 HV Amsterdam
E-mail	antske.fokkens@vu.nl
Tel	+31 20 59 877 51
CLARIAH Component	wp3-robust-semantic-parsing-Dutch (in particular, simple-tagger with HISCO, NLP2RDF BiographyNet)

Name	Auke Rijpma
Function	Post-doc researcher
Organization	Utrecht University
Address	Drift 6 / 3512 BS / Utrecht

E-mail	a.rijpma@uu.nl
Tel	+31 30 253 6460
CLARIAH Component	wp4-converters, specifically HISCO and HISCAM

Name	Martijn Kleppe
Function	Researcher
Organization	National Library of the Netherlands (KB)
Address	Prins Willem-Alexanderhof 5, 2595 BE, Den Haag
E-mail	martijn.kleppe@kb.nl
Tel	070-3140911
CLARIAH Component	Not applicable

4. Centre

Centre	Huygens ING
Name	Jauco Noordzij
Function	Architect and Lead Engineer of WP-2
Address	Oudezijds Achterburgwal 185, 1012 DK, Amsterdam
E-mail	jauco.noordzij@huygens.knaw.nl
Tel	+31 20 224 685 2

The Huygens will incorporate the generated results in their database. We use (and improve) existing CLARIAH tools. The tooling created in this project will thus be taken up in existing CLARIAH structures.

5. Description of the Proposed Project

5.1 Research Question(s)

Around the turn of the century, post-war trends of declining inequality halted or even reversed in many countries. The public concern over this issue has become evident with the media attention of academic work on trends of economic and social inequality (e.g. Pikkety and Ganser 2014, Clark 2014, Van Zanden et al. 2014).

This proposal meets the public concern and academic interest by evaluating whether Robust Semantic Parsing tools can be used to answer historically substantive questions on social inequality, namely:

1. How did career patterns in the Netherlands in the long nineteenth century look and how did they change over time?
2. To what extent do job advertisements in Dutch news papers reflect the hypothesized change from ascription to achievement for occupational attainment?

Both questions have recently been studied (e.g. Schulz & Maas 2010, Schulz, Maas & Van Leeuwen, 2015). Our aim is to ‘replicate’ those studies using innovative methods. Previous work focuses on linking events (birth, marriage and death records or census data) or base conclusions on small samples of biographies. Linkage approaches are limited because many variables introduce linkage biases. This does not apply to biographies, but their manual analysis is time consuming. We circumvent this through Robust Semantic Parsing Tools (WP-3) in combination with Linked Data converters (WP-4) to extract career paths from approximately 35,000 biographies from the Biography Portal (BP;

WP-2). We furthermore extract information on the ascribed and achieved characteristics asked in job advertisements in Dutch newspapers from the Royal Library (KB), trying to automate the approach taken by Schulz (2015) who annotated more than 2000 job advertisements manually. Job advertisements introduce a methodological challenge in the quality of their OCR. We address the following methodological questions:

- How capable are CLARIAH tools to automatically extract information on occupations and occupational characteristics from qualitative sources, such as biographies and job advertisements?
- To what extent is the performance of CLARIAH semantic processing tools hampered by OCR quality?

We addresses the period of 1815-1940 which is covered by all corpora (the BP, KB and the Historical Sample of the Netherlands (HSN)).

5.2 CLARIAH Component(s)

We use CLARIAH components from three work packages.

WP2: Person Data

The Person Data from the BP are more likely to reflect the upper part of society, but we are not aware of systematic research on the social background of persons in the BP. Most BP biographies include information about the subject's career, education and the (grand)parents's occupations. By comparing social background and educational attainment with existing studies based on HSN (e.g. Zijdeman, 2010), we can assess the social selectivity of this important source and changes therein over time.

WP3: Dutch Robust Semantic Parsing

We use the **Dutch Robust Semantic Parsing Tools** for text mining, notably:

- Word Sense Disambiguation (WSD)
- The simple tagger with occupation linking (HISCO tagger)
- NLP2RDF crystallization (the BiographyNet variation)

The HISCO tagger identifies occupations mentioned in text and links them to their Hisco code. WSD resolves ambiguities (e.g. *broeder* can mean ``monk'' or ``brother''). The BiographyNet NLP2RDF crystallization links occupations to people (Fokkens et al. forthcoming).

CLARIAH work on interoperability allows us to incorporate tools using FoLiA, such as Reynaert's (2014) tool for cleaning OCRed text.

WP4: converters for HISCO and HISCAM

In combination with WP-4's *data-converters for HISCO* (Maas, van Leeuwen & Miles 2004) *and HISCAM* (Lambert et al. 2013), we can systematically process occupation information in an internationally and temporally comparative manner. HISCO allows us to classify the occupations based on activities of the incumbents of those occupations. Within CLARIAH, the classification has been converted to RDF. HISCAM ranks occupations according to the occupational stratification structure of the long nineteenth century.

5.3 Description

How CLARIAH components will be used

We apply the semantic parsing tools to biographical text and job advertisements. These tools identify among others *named entities, semantic roles, word senses, mentions of professions, temporal expressions* and link professions and family relations to people. This structured data can be used to extract profession patterns.

Expected Outcome

Our main result is an evaluation of new possibilities to use qualitative sources in an automated way. Methodologically, we evaluate the extent to which CLARIAH tools are up to this task, and improve upon the tools to circumvent the expected challenges mentioned below. Specifically:

1. Addition of extracted career patterns to the BP
2. Comparison of ascriptive and achievement requests of large volumes of job advertisements
3. Improved HISCO tagger addressing challenges below
4. CLARIAH tools:
    a. Improvement *semantic parsing tools* by combining with alternative tools
    b. Extensive evaluation reports (error analysis and plans)

Challenges and Solutions:

We anticipate two main challenges:

1. The tools we are using are mainly developed for and trained on contemporary Dutch.
    a. Lower performance on old Dutch.
    b. Lower performance on job advertisements due to OCR quality.
2. The performance of the simple-tagger in both recall and precision:
    a. The HISCO tagger is string-based and erroneously links ambiguous terms to occupations (e.g. *broeder* to `monk’ instead of `brother’). We address this by applying word sense disambiguation.
    b. The linguistic expressions are derived from the extensive occupations list of the HSN (Mandemakers et al. 2013), which was automatically cleaned. Lack of coverage and errors in cleaning are expected. We address this by semi-automatically cleaning existing expressions and mining new ones.

5.4 Plan

The research is carried out in three cycles of four months. During each cycle, we process the data with the current version of semantic parsing tools. The outcome is evaluated against gold data available at that time. The social historians investigate the answers to their research questions using the output. We use this investigation in cycle 1 to select additional evaluation data for cycle 2. This evaluation data should include enough data that supports the main trend of the outcome, enough that contradicts it (the exceptions) and enough neutral data. This increases the chances of spotting errors in the automatic analysis that introduce a bias influencing the overall results.
In our evaluation, we stay methodologically close to related work (Schultz et al. 2015; Zijdeman 2009). This means we apply multilevel growth models to the HISCAM-career trajectories. Considerable attention goes to the validation of the data and the new approach, we may limit the multilevel aspect to time and place fixed effects. During each cycle, we evaluate whether the results fit the social mobility literature.

Current state of affairs
The *Semantic Parsing tools* contain all required tools (with different levels of accuracy).
Expected begin state
Work planned independently of this project (by March): ● Automatic cleaning steps to improve HSN expressions in HISCO tagger ● Add expressions from Brouwers to HISCO tagger ● Evaluation named entities, coreference on biographical data
Cycle 1 (1 May - 31 Aug)

- Add occupation and family relations to evaluation set biographical data
- Manual check of automatically cleaned expressions
- Apply *semantic parsing* to biographies and advertisements
- Design first improvements of HISCO tagger:
    - Disambiguate terms
    - Identify occupations not mentioned in the HSN list
- Analyze outcome:
    - Evaluation: which tools need improvement?
    - Sample evaluation data:
        - from various periods
        - with various degrees of relevant information

Cycle 2 (1 September - 31 December)

- Annotate sampled data
- Integrate alternative tools (e.g. LaMachine)
- Implement design cycle 1
- Evaluate alternative versions of tools

Cycle 3 (1 January - 30 April)

- Incorporate minor improvements based on evaluation
- Document tools and evaluation
- Integrate outcome into Person Data Huygens
- Write papers

6. Deliverables and Milestones

Id	Title	Type	Description	Responsible	Planned
1.	BP Evaluation	Dataset	A balanced set from the Biography Portal annotated with career information (occupation, education).	Zijdeman	Month 4
2.	KB-JA Evaluation	Dataset	A set of annotated job advertisements from the KB. Annotated for ascribed and achieved characteristics	Rijpma	Month 8
3.	Wishlist	Doc	Outline of desired improvements of CLARIAH tools (two versions)	Zijdeman	Month 4 & 8
4.	HISCO tagger	Software	Improved version of HISCO profession tagger; augmented with additional expressions and variations	Zijdeman	Month 8
5.	Tagger doc.	Doc	Documentation on the improvements and current status of the HISCO tagger	Rijpma	Month 10
6.	BN RDF data	Dataset	Enhanced/improved enrichments for the Biography Portal	Fokkens	Month 9

7.	BN RDF report	Doc	Document describing the generated results	Rijpma	Month 10
8.	semparse evaluation	Doc	Report on CLARIAH tool evaluation; detailed error analysis of *semantic parsing tools,* improvements during the project and wishlist for future work	Fokkens	Month 12
9.	SH paper	Doc	Paper describing the outcome of the career analysis. Target (impact allowing): Historical Life Course Studies (European Historical Population Samples network journal).	Rijpma	Month 12
10.	DH paper	Doc	Paper describing the methodological aspects of the study. Target (impact allowing): journal of digital humanities.	Fokkens	Month 12
11.	Project report	Doc	Report describing research questions, the CLARIAH tools used and detailed evaluation.	Zijdeman	Month 12

7. IPR and Ethical Issues: Risks

Ethical Issues are not applicable for the Biography Portal since this only includes people who have deceased. Different parts of the Portal have different licenses, but all data is available for researchers at Huygens and partners. Extracted patterns can be made publically available. The KB Newspaper data is copyright protected but will be made available to the research team for internal use via the KB Dataservice, since HHu'Cap will not republish the articles but use it as research material.[1]

8. Expertise of the applicant(s)

Antske Fokkens is a computational linguist who has been active in the field of digital humanities since 2012. She was responsible for the text mining components in BiographyNet and worked on data modeling and interoperability within the NewsReader project. She has designed several of the tools used in this project and worked extensively with the others. She currently works on her VENI project for 4 days a week and spends 1 day working on CLARIAH (mainly interoperability). Her research of the past years focuses on the methodology of text mining for digital humanities. She will be responsible for the text mining component which will be carried out in collaboration with other researchers at the Vrije Universiteit. She is currently leading a team of student assistants that is annotating biographies and will co-supervise the student assistants annotating the evaluation data.

Martijn Kleppe works at the Research Department of the National Library of the Netherlands (KB). He is advisor Digital Scholarships and works on several Digital Humanities projects. Before he worked at the Erasmus University Rotterdam and Vrije Universiteit Amsterdam. He led DH projects PoliMedia (www.polimedia.nl), Talk of Europe (www.talkofeurope.eu) and was work package leader of the Fp7-project AXES-Access to Audiovisual Archives (www.axes-project.eu). He is interested in enhancing the potential of the KB data and will have a, to the project costless, consultancy role in this project.

Jauco Noordzij is the lead engineer of the team that is responsible for most of the WP2 infrastructure components. His experience lies in software engineering practices and software release management. He is responsible for the RDF representation of the Biography Portal. He is currently collaborating with Niels Ockeloen and Antske Fokkens to

[1] https://www.kb.nl/bronnen-zoekwijzers/dataservices-en-apis

incorporate the results of the BiographyNet project in the Biography Portal, where focus lies on marking the provenance of information. He will be responsible for adding the newly generated data to this repository. Clear marking of provenance will be essential in this step as well.

Auke Rijpma is the data manager for the Clariah Structured Data Hub (wp4) and assistant professor in economic and social history at Utrecht University. He approaches the topic from a data-oriented, quantitative perspective and has worked on diverse topics such as the long term measurement of wellbeing, human capital and fertility, gender equality, social spending, and medieval economic development. He will partake in the social-historical investigation of the project and help with the evaluation of the quality of the data.

Richard Zijdeman is project leader for the Clariah Structured Data Hub (wp-4) and senior researcher and chief data officer at the IISH. His substantive research expertise lies in the field of long term occupational stratification and social mobility with specific focus on the influence of technical innovations (e.g. industrialization, mass transportation) in the long Nineteenth century. Zijdeman's methodological focus is on applying hierarchical models on clustered data and development of tools for occupational coding and measures of occupational stratification. He will coordinate the social-historical investigation of the project and co-supervise student assistants creating evaluation data and in their efforts of cleaning the linguistic component of the HISCO tagger.

9. Reviewers

Suggested reviewers: (name, affiliation, e-mail address)

Prof. dr. Paul S. Lambert, Stirling University, paul.lambert@stir.ac.uk
Dr. Marten Düring, University of Luxembourg, marten.during@uni.lu

10. Project budget details

Participant	Organization	Effort (PM)	Salary Costs/PM (Euro)	Salary Costs (Euro)
Richard Zijdeman	IISG	2	€ 6.265,00	€ 12.530,00
student assistent 1	IISG / VU	2	€3.072,00	€ 6.144,00
student assistent 2	IISG /VU	2	€3.072,00	€6.144,00
Antske Fokkens	VU	3	€6.265,00	€18.795,00
Auke Rijpma	Utrecht Uni	1.5	€ 6.265,00	€ 9.397,50
Jauco Noordzij	ING/Huygens	1	€ 6.265,00	€ 6.265,00
Subtotal				**€ 59.275,50**

Material Costs

Item	Organisation	Costs/ Unit	#Units	Costs
Travel + Posters	all	€ 250,00	1	€ 250,00
Subtotal				**€ 250,00**
Total				**€ 59.525,50**

11. Literature

Clark, G. 2014. *The son also rises: surnames and the history of social mobility*. Princeton University Press.

Fokkens, A.S., S. ter Braake, N. Ockeloen, P. Vossen, S. Legêne, G. Schreiber and V. de Boer. forthcoming.*BiographyNet: Extracting Relations between People and Events*.

Lambert, P.S., R.L. Zijdeman, M.H.D. Van Leeuwen, I.Maas, and K.Prandy. 2013. “The Construction of HISCAM: A Stratification Scale Based on Social Interactions for Historical Comparative Research.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 46 (2): 77–89. DOI:10.1080/01615440.2012.715569

Van Leeuwen, M.H.D., I. Maas and A. Miles. 2004. Creating an Historical International Standard Classification of Occupations (HISCO). An exercise in Multinational Interdisciplinary Co-Operation. *Historical Methods*, 37, pp. 186-197.

Mandemakers, K., S. Muurling, I. Maas, B. van de Putte, R.L. Zijdeman, P.S. Lambert, M.H.D. van Leeuwen, F. van Poppel and A. Miles. 2013. *HSN standardized, HISCO-coded and classified occupational titles, release 2013.01*, Amsterdam:IISG.

Piketty, T., and L. J. Ganser. 2014. *Capital in the twenty-first century*. Belknap press.

Reynaert, M.W.C. 2014. Synergy of Nederlab and @PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up. In: Proceedings of LREC. Reykjavik, Iceland: ELRA.

Schulz, W. and I. Maas. 2010. Studying historical occupational careers with multilevel growth models. Demographic Research 23 (24), 669-696. DOI: 10.4054/DemRes.2010.23.24

Schulz, W., I. Maas, and M.H.D. van Leeuwen. 2015. “Occupational Career Attainment during Modernization. A Study of Dutch Men in 841 Municipalities between 1865 and 1928.” *Acta Sociologica* 58 (1): 5–24. DOI:10.1177/0001699314565795.

van Zanden, JL., J. Baten, M. Mira D’Ercole, A. Rijpma, C. Smith, and M. Timmer, eds. 2014. *How Was Life? Global Well-Being since 1820*. Paris: OECD Publishing. DOI: 10.1787/9789264214262-en.

Zijdeman, R. L. 2009. “Like My Father before Me: Intergenerational Occupational Status Transfer during Industrialization (Zeeland, 1811–1915).” *Continuity and Change* 24 (3): 455–486.

KB

Koninklijke Bibliotheek

Prins Willem-Alexanderhof 5
Postbus 90407
2509 LK Den Haag

Telefoon
(070) 314 09 11

Fax
(070) 314 04 50

Website
www.kb.nl

CLARIAH Office
Attn. Selection Committee Research
Pilots Call
Trans 10
3512 JK Utrecht

Datum

3 November 2016

Betreft

Support letter research project
History of Human Capital

Direct nummer

+31703140414

E-mail

Hans.jansen@kb.nl

Dear members of the selection committee of the CLARIAH Research Pilots call,

We would like to inform you that The National Library of the Netherlands (KB) is enthusiastic about the ‘History of Human Capital (HHu’Cap)’ proposal and we would like to express our support for it.

We will give the research team full access to our newspaper collections and are eager to follow the teams progress in applying CLARIAHs tools to investigate the possibilities of extracting personal information from our newspaper articles that stem from our collection.

We look forward in our support to make HHu’Cap a successful research project.

With kind regards,

Hans Jansen
Head Research Department