Towards an Open Science Infrastructure for the Digital Humanities: The Case of CLARIN

Koenraad De Smedt[1], Franciska de Jong[2], Bente Maegaard[3], Darja Fišer[4] and Dieter Van Uytvanck[5]

[1] University of Bergen, Norway
[2,5] CLARIN ERIC, The Netherlands
[3] University of Copenhagen, Denmark
[4] University of Ljubljana and Jožef Stefan Institute, Slovenia
clarin@clarin.eu

Abstract. CLARIN is the European research infrastructure for language resources. It is a sustainable home for digital research data in the humanities and it also offers tools and services for annotation, analysis and modeling. The scope and structure of CLARIN enable a wide range of studies and approaches, including comparative studies across regions, periods, languages and cultures. CLARIN does not see itself as a stand-alone facility, but rather as a player in making the vision that is underlying the emerging European policies towards Open Science a reality, by interconnecting researchers across national and discipline borders and by offering seamless access to data and services in line with the FAIR data principles. CLARIN also aims to contribute to responsible data science by the design as well as the governance of its infrastructure and to achieve an appropriate and transparent division of responsibilities between data providers, technical centres, and end users. CLARIN offers training towards digital scholarship for humanities scholars and aims at increased uptake from this audience.

Keywords: CLARIN, Research Infrastructure, Language Resources and Technologies.

1 Introduction

CLARIN, the European research infrastructure for language resources, provides access to digital language resources and tools through a single sign-on environment with the aim to support researchers in the humanities and social sciences and related fields. In 2012 it was established as a European Research Infrastructure Consortium (ERIC), for which basic funding comes from the member countries; since then it has grown from nine members to nineteen members and two observers.[1] This growth, which is still on-going, shows that the concept underlying CLARIN was valid and that the start of the implementation was timely, not only from the perspective of the state-of-the-art of the infrastructural technology, but also in view of the demand for

[1] Additionally, CLARIN has a special agreement with Carnegie Mellon University in the USA.

support from the target domains. More specifically the rapid increase of interest in digital scholarship in the humanities has reinforced the potential for impact.

While the digital humanities (DH) have been evolving from existential debates (e.g. Burdick et al. 2012; Gold 2012; Svensson 2009; Terras et al. 2013) to a more confidently outspoken community of practice (e.g. Schreibman et al. 2016), CLARIN has been taking a pragmatic approach to what the DH need in terms of research infrastructure. It has been pointed out by DH scholars that "Curation, analysis, editing, and modeling comprise fundamental activities at the core of DH. Involving archives, collections, repositories, and other aggregations of materials, curation is the selection and organization of materials in an interpretive framework, argument, or exhibit. The capacity with digital media to create enhanced forms of curation brings humanistic values into play in ways that were difficult to achieve in traditional museum or library settings." (Burdick et al. 2012:17–18). Organized support for digital curation, analysis, editing and modeling involves "platforms, tools, and infrastructures" which "depend upon the basic building blocks of digital activity: digitization, classification, description and metadata, organization, and navigation" (Burdick et al. 2012:17).

It is in this setting that CLARIN operates, not by forcing a model on the DH or institutionalizing it, but by contributing an infrastructure and meeting ground which aims to make "all digital language resources and tools from all over Europe and beyond [...] accessible [...] for the support of researchers in the humanities and social sciences" (Maegaard et al. 2017:2).

2 Scope and Relevance for DH

2.1 Scope of CLARIN

An important factor for the success and sustainability of a research infrastructure such as CLARIN is its scope, size and structure. CLARIN deals with digital language data and their curation and processing. The observation that "[t]ext encoding seems to create the foundation for almost any use of computers in the humanities" (De Smedt 2002:95) largely still holds. Language is an essential instrument for human cognition and expression, a rich carrier of cultural content, a reflection of societal dynamics, and a central part of the identity of individuals and groups. Language materials, in all their forms, synchronically and diachronically, are therefore a core object of study in the humanities.

The digitization of language materials in traditional infrastructures for humanities scholarship, such as libraries, archives and museums, as well as the continuous creation of new digital research data on the computers of humanities scholars, bring about the need for new forms of organizing and sharing these materials so as to promote their findability, accessibility, citeability and permanence, and to offer tools and services for their analysis.

CLARIN is collecting data for languages from all places and periods that are of interest to the European research area and beyond and integrates all metadata on its

central platform. Consequently, CLARIN offers support for the study of national and regional languages and cultures, but in addition, it is the combination of multiple resources and the analytic tools available for multiple languages that makes CLARIN an enabler of comparative studies across regions, periods, languages and cultures. This support extends to the study of phenomena that are characteristic for the culture of Europe based on language data, such as language variation, multilinguality, migration patterns, intellectual history, etc.

Furthermore, CLARIN has recently started targeted actions that promote the findability and visibility of specific data types and families of resources that are relevant for humanist research agendas; so far these actions have focused on users of four types of materials: newspapers, oral history data, parliamentary data and social media data, with more to come (Fišer et al., 2018).

CLARIN has established the Virtual Language Observatory (VLO)[2], a registry of Language Resources (LRs) based on the CMDI metadata standard. The VLO contains information about all LRs made available in the member countries, plus information from other registries that want to be visible through the VLO (Van Uytvanck et al. 2012).

CLARIN is now aiming at more cross-institutional and cross-sectorial collaboration, e.g., with the GLAM sector (Galleries, Libraries, Archives, Museums) and with industry. Collaboration with other research infrastructures, be it in the humanities and social sciences area or with eScience, is also pursued in order to foster multidisciplinarity and the inherent need for the innovation of methodological frameworks. Through cross-border collaboration, as well as through the focus on training and education, on centres of expertise etc. CLARIN will increase its societal impact and will contribute to the development of methodologies for measuring such impact. Finally, an inherent goal of all these activities is to integrate and contribute to Europe's Open Science policies[3] as will be discussed in Section 3.

2.2 **Research Questions as Driving Forces**

The CLARIN infrastructure supports researchers in identifying relevant data and tools, and contributes to the reuse of data created by scholars. In the various national consortia, CLARIN has spawned an ecosystem of collaborative projects involving a wide range of disciplinary communities, including literary studies, history, political studies, philology, history, media studies, corpus linguistics, etc. In this subsection we will describe some examples of datasets and collections that have been generated with the aim to strengthen the basis for comparative research, enabled by the consultation or participation of humanities scholars for the work on the curation and annotation of data. In order for the resources to contribute to the advancement of the kind of insights that humanities scholars are after, a crucial condition is that the resources have added value in answering a research question or in helping to shape a research agenda. We will therefore briefly mention a few examples of humanities research

[2] http://vlo.clarin.eu
[3] http://ec.europa.eu/research/openscience/

questions for which resources, methods and tools available through CLARIN have proven instrumental.

It should be underlined that for these (and other) cases the collaboration with domain experts is equally important. See Kestemont et al. (2017) for an illustration of how in the case of solving the riddle of the authorship of the Dutch national anthem, alternative hypotheses could be suggested based on text mining, while only the confrontation with existing insights in the 16th century cultural context and the literary conventions at the time of creation could bring the authorship attribution to a next level of validation.

The work done in the context of the Talk of Europe project[4] on parliamentary data is a typical example of data enrichment, paving the way for a multidisciplinary agenda. Parliamentary recordings, and texts, such as transcribed debates and speeches, are of relevance for studying, for instance, how historical, cultural and religious attitudes are reflected in political discourse. Van Aggelen et al. (2016) established LinkedEP, a Linked Open Data translation of the verbatim reports of the plenary meetings of the European Parliament and enriched by links to other data. Scholars have used this dataset to study terms over time, for example in examining how the financial crisis was discussed in the European Parliament. The language captured in parliamentary records can also be studied as a carrier of emotion, and of the correlation with other phenomena related to cultural and or social dynamics (e.g. Rheault et al. 2017).

Another use case in the domain of political studies is the work by Andreas Blätte, who explores possibilities to combine an interpretative approach to analyse the discourses constructing policy fields with quantifications of textual data. For the purpose of entity extraction, CLARIN tools were applied to corpora of plenary debates for the German Bundestag and the regional parliaments (Blätte and Blessing, forthcoming). One of the results has been an analysis and visualisation of word relations in the discourse on the politics of integration across political parties and periods.[5]

Similar benefits of tools for processing larger datasets have been reported by researchers that deploy topic modelling and visualisation tools. Martinez-Ortiz et al. (2016), for example, applied this to the question: Which concepts of 'war' do newspapers reveal? Their comparative analysis of Dutch newspapers over time and space (presented at the CLARIN-PLUS workshop on working with digital collections of newspapers) was supported by tools that generate graphs like that in Fig. 1, showing that the term *oorlog* ('war') has been a shifting concept.

[4] Funded by CLARIN ERIC and CLARIN-NL.

[5] The quantitative analysis tool set has been made available under the name PolMineR at http://www.github.com/PolMine/polmineR.

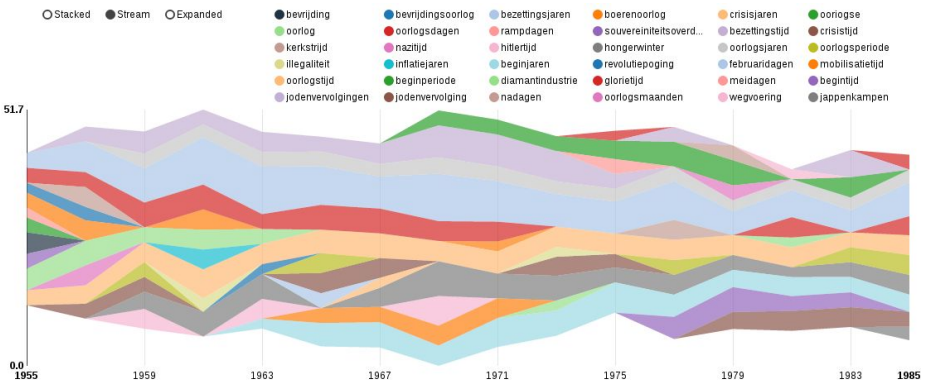

Fig. 1. Visualization of time shifting vocabulary related to *oorlog* ('war') in Dutch newspapers (reproduced from Martinez-Ortiz et al. 2016). Layer height represents word frequency.

An example of a study integrating insights on cultural dynamics and intellectual history is the work by Goldhahn et al. (2017), who investigate how Ernst Jünger's nationalistic vocabulary can be seen in a temporal dimension and in the perspective of contemporary newspaper language. The authors used tools developed in the German CLARIN-D, including Corpus Diff, to compare and visualize the various sources, and WebLicht, which supports the selection and execution of tool chains for text processing without the need of downloading or installing any software (Hinrichs et al., 2010). Tiepmar et al. (2017) have taken such tool development further by integrating a Canonical Text Service in CLARIN; this allows the comparison of text editions, supporting diachronic and synchronic research on textual variation, through an interface which is intuitive for humanities scholars.

Several projects supported by the earlier Dutch CLARIN-NL used CLARIN data and tools to support innovative digital research in the humanities, including the following examples. Correspondence patterns and learned practices in the 17th century Dutch Republic were established using adapted linguistic analysis tools on a database of 20,020 letters in TEI-format (Ravenek et al. 2017). A DH approach to the history of culturure and science, focusing on drugs and eugenics in early 20th century Dutch newspapers was carried out through semantic text mining (Snelders et al. 2017). A digital workbench for research into the life and works of Rembrandt was created using the federated search infrastructure (Verberne et al. 2017). The landscape of names in Modern Dutch Literature was mapped using tools for named entity recognition on a TEI-encoded corpus, a semanticiser tool which tries to link named entities in the texts to entries in Wikipedia, and a visualiser (De Does et al. 2017).

The growing number of digital interview and oral history collections that are accessible through CLARIN has given rise to studies into how eyewitnesses have become ever more prominent in the media and how a better insight could enhance our understanding of the different functions and values attributed to testimonies and of how individuals recall mass violence. In the context of the current Dutch CLARIAH (with relations to both CLARIN and DARIAH), a cross-media and diachronic content analysis is being conducted of eyewitness testimonies (EWTs) in newspapers, on

radio and television, and in oral history interviews in the Netherlands since 1945.[6] The proposed interview transcription chain is shown in Fig. 2.

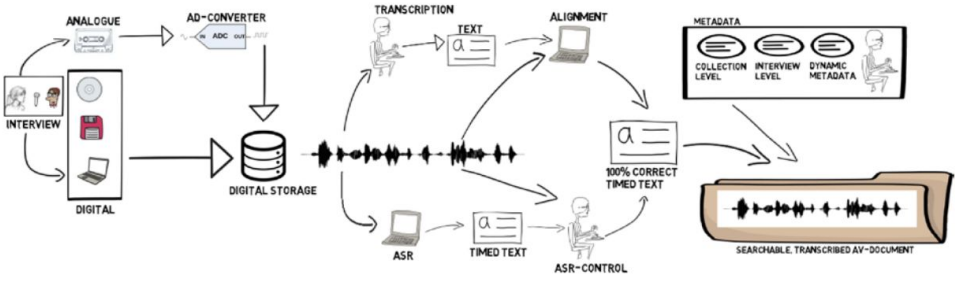

Fig. 2. Proposed EWT transcription chain (CLARIAH project presented at the CLARIN workshop on Oral History).

3 CLARIN as an Open Science Infrastructure

Since "the Digital Humanities remains at its core a profoundly collaborative enterprise" (Burdick et al. 2012:ix), common methodologies should be available for all researchers to draw on. DH transcends discipline boundaries through a methodological commons where data and tools are, to the largest possible extent, accessible for all to use in a spirit of Open Science. "CLARIN does not see itself as a stand-alone facility, but rather as a player in making the vision that is underlying the emerging European policies towards Open Science a reality, interconnecting researchers across national and discipline borders by offering seamless access to data and services in line with the FAIR data principles." (Maegaard et al. 2017:3; cf. also De Jong et al. 2018). The FAIR data principles[7] are now widely promoted as part of the Open Science paradigm (Wilkinson et al. 2016). In the paragraphs below we outline the network architecture of CLARIN and demonstrate how it is FAIR-compliant.

CLARIN comprises not only central services like the VLO that give access to what is on offer in the distributed collection of resources, but also consists of a network of more than 40 certified centres which provide data and services for their curation, analysis, modeling and knowledge sharing. A CLARIN centre typically provides a data repository which offers a sustainable home to documented research data which may be the output of projects, research groups or individual scholars. Additionally, many centres also provide tools (web applications, web services or stand-alone applications) to process language data. For an overview of the CLARIN technical infrastructure, see Odijk (2017).

[6] https://www.clariah.nl/projecten/research-pilots/crossewt

[7] https://www.force11.org/group/fairgroup/fairprinciples

3.1 Findable, Accessible, Interoperable, Re-usable

Open access to language resources can only be realized when they can be *found* and reliably *identified* by researchers. FAIR therefore requires persistent identifiers (PIDs), which secure the citeability of data, and rich metadata which is indexed and searchable, with links from the metadata to the data identifier. CLARIN requires CMDI metadata (Goosen et al. 2015) to describe the data and tools in repositories; these metadata are open and are harvested from centres into catalogues (including the VLO) which make them findable.

Language data should be easy to *access*. FAIR translates this into the need for a standardized communication protocol, with the option for easy authentication and authorisation when needed. CLARIN relies on the HTTP protocol and SAML for federated single sign-on. Resources are as open as possible, but whenever restrictions are necessary (e.g., due to privacy or copyright considerations) the conditions for use are made explicit.

To attain *interoperability*, FAIR demands a formal, shared and broadly applicable language for knowledge representation, using FAIR vocabularies and links between metadata and data. For its metadata CLARIN relies on the CMDI framework as common metadata language, including links to standardized OpenSKOS vocabularies (e.g. Brugman 2017) and standardized ways of linking to datasets and landing pages. There are also recommendations for the use of standard data formats, such as TEI.

Finally, FAIR states that *re-usable* data requires clear license and provenance information and adherence to community standards. CLARIN has clear recommendations on license disclosing and user-friendly ways of categorizing these (Arppe et al. 2011). The provenance needs to be part of the metadata. While community standards are hard to define, the bottom-up structure of the centres definitely brings along close ties with such good practices.

3.2 Usability and Training towards Digital Scholarship for Humanists

With the wide emergence of digitally available language data (be it digitally native or digitized analogue resources), the possibilities beyond the mere archiving and viewing of such data sets have significantly grown. However, DH is sometimes perceived as ‘inaccessible’: while the tech-savvy researchers use ground-breaking methodologies, the majority of scholars read but do not participate, prompting this question: “are the barriers to entry that ‘outsiders’ perceive real usability issues, or simply points on DH’s learning curve?” (Edwards 2012, p. 213).

Why not address both potential bottlenecks? On the one hand, CLARIN explicitly addresses the usability of its data and methods, not only through cataloguing and adequate documentation, but also through a new facility called the Language Resource Switchboard, which supports researchers in identifying the tools that are suitable for use with the given data types and to decide on a workflow that fits the research question they want to address (Zinn 2016). Thus, CLARIN aims to provide flexible connections between data, tools, and standards, never forcing particular

models or methods, but seamlessly supporting the productive (cf. Edmond 2016:60). Furthermore, CLARIN centres are offering a range of online tools for search (including Federated Content Search, which simultaneously searches several nodes in the network), tools for analysis, and workflow systems, which allows researchers to tailor generic scenarios to a variety of user perspectives. Nevertheless, there is a continuous need for updating user interfaces to improve the user experience for current and new audiences.

On the other hand, the learning curve is made manageable through extensive knowledge sharing and user involvement. CLARIN operates a Knowledge Sharing Infrastructure (KSI) which employs several instruments in order to share knowledge of digital methodologies in the humanities. CLARIN has nine knowledge centres at the time of writing, and the number is still expanding. Knowledge centres may focus on the languages of a country, or on a specific technology or type of data (e.g. audio-visual data) and offer expertise to researchers. Uptake is promoted through live events (such as dedicated summer schools,[8] researcher training courses, tutorials at conferences, master classes, etc.) and on-line training materials, which demonstrate the functionalities of the available materials and tools to newcomer audiences. Workshops are key to sharing new knowledge and developing the methodological apparatus in specific areas, e.g., the workshop on exploring spoken word data in Oral History archives (2016),[9] or the workshop on working with parliamentary records (2017).[10] Workshops may be used to support new user communities, to make the needs of a particular community visible and to work towards the advancement of methodologies across countries as well as disciplinary borders.

4 Responsible Data Science

Across almost all domains of research there is growing concern for how data, especially 'big data', are put to use. DH uses big data approaches, for instance, in 'distant reading' (Moretti 2005, 2013), a method which "engages the abilities of natural language processing to extract the gist of a whole mass of texts and summarize them for a human reader in ways that allow researchers to detect large-scale trends, patterns, and relationships that are not discernable from a single text or detailed analysis" (Burdick et al. 2012:39). Even if such methods unleash the power of digital scholarship on data that are simply too large to read, critical reflection is mandatory, both with respect to the methods that filter and extract patterns in ways that can be opaque, and the parameters which are used to collect and annotate the data in the first place.

Any use of data which is biased, violates privacy or confidentiality, or lacks transparency, may distort conclusions or break trust relations. A recent expression of

[8] For example, the annual European Summer University in Digital Humanities in Leipzig: http://www.culingtec.uni-leipzig.de/ESU_C_T/node/97

[9] https://www.clarin.eu/event/2016/clarin-plus-workshop-exploring-spoken-word-data-oral-history-archives

[10] https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records

these concerns is by the Responsible Data Science consortium (RDS)[11] which aims to tackle ethical and legal challenges, promote data science techniques, infrastructures and approaches that are responsible in the sense that data and data use should be fair, accurate, confidential and transparent (FACT) (Van der Aalst et al. 2017). These aims complement the FAIR principles, especially in a context where the use of data-driven methods typically applied to larger datasets is on the rise.

The concerns in FACT also pertain to language data in the humanities and social sciences. CLARIN is not primarily geared toward big data, but rather toward quality data; it intends to contribute to responsible data science by the design as well as the governance of its infrastructure and to achieve an appropriate and transparent division of responsibilities between data providers, technical centres, and end users. Data curation is a core task for CLARIN data centres (repositories). License agreements, which are established between a data provider and a data centre, regulate the terms under which some well-described data is made available. These terms include an end-user license agreement which, together with the terms of service at the data centre, may place some restrictions and responsibilities on the end user, particularly in the case of privacy concerns. The requirement of provenance data in CLARIN metadata makes data traceable and the use of PIDs makes data citable and their use replicable. Furthermore, CLARIN has started to provide guidance on which tool is recommended for which data through the above-mentioned Language Resource Switchboard. Plans to enhance CLARIN data include improved sampling, enrichment of the data with (extralinguistic) metadata, as well as linking with external knowledge sources (e.g. gazetteers) and annotation at the conceptual level.

Additional steps are foreseen which relate to the need to document and explain the performance levels that can be expected from the analysis tools, and thereby of the suitability of certain tools for specific scenarios of use. Scenario-based testing is particularly relevant for the uptake of CLARIN functionality in the context of multidisciplinary collaboration where methodological frameworks rooted in humanities traditions will have to be combined with what has roots in other scholarly traditions.

The step from big data to big conclusions and decisions requires a considerable level of transparency from the algorithms applied, since black box applications are not likely to be accepted as the basis for conclusions. This issue is addressed in discussions and workshops addressing the concept of *tool criticism*, that can be seen as complementing the humanist tradition of source criticism (Van Ossenbruggen, 2017). Sustainable scientific and societal impact of the tools on offer can only be expected if the validity of analysis results can be explained to and assessed by relevant researcher communities (Manovich 2016; Nguyen et al. 2016). In conclusion, infrastructures cannot and should not take over the responsibility for assessing the appropriateness of data and methods, but at least infrastructures can provide researchers with information that empowers them to carry out such assessments.

[11] http://www.responsibledatascience.org

5 Concluding Remarks

As DH researchers "invest at an unprecedented level in the essential substrate of their research" (Edmond 2016:63), CLARIN is helping to build, secure and exploit this investment. From its inception more than 10 years ago, CLARIN has played a role in digital transformations in the humanities (cf. Wynne 2013), starting at a time when DH had not yet gathered the steam it has today. Since then, CLARIN has become a potent player that interconnects researchers across borders by making digital data and tools more accessible.

In this paper we have indicated the relevance of CLARIN for an increasing number of disciplines and approaches in the humanities and social sciences, in particular for DH. We have also shown how, from the outset, CLARIN has been in line with the principles for FAIR data that have recently been made explicit. Plans for being compliant with the Responsible Data Science framework are being developed, as well as efforts to reinforce the multidisciplinary potential of CLARIN.

The FAIR principles should not be taken as ukazes, but as helping hands that enhance the visibility and citeability of DH project results. A way towards this goal would be the promotion of common Data Management Plans (DMPs) and related information platforms for humanists (e.g. Trippel and Zinn 2016). In addition a framework for domain-specific data protocols is being developed by an alliance of stakeholders (Science Europe 2018). CLARIN has also been participating in several other European projects aimed at sustaining, consolidating, enhancing and widening research infrastructure: CLARIN-PLUS, EUDAT2020, Language Technology Observatory, EUROPEANA-DSI and EOSC-hub.

The ultimate impact of CLARIN will be in its uptake by researchers and its relevance for stakeholders both inside and outside of academia. Current and planned CLARIN efforts towards uptake reflect and address the CLARIN vision with multiple strands of core activities. A series of surveys have been designed to evaluate the comprehensiveness and usability of CLARIN services and prioritise future development efforts. In parallel, training models have been developed that stimulate the uptake of CLARIN resources, tools and services by researchers from a wide range of disciplines in the humanities and social sciences. Furthermore, two focus groups with researchers using the CLARIN have been carried out, in which participants share their experiences and problems with the infrastructure as well as their needs and suggestions. One group involved researchers with a strong technical background, such as Natural Language Processing or Text Mining, and one involving humanities and social sciences researchers with limited technical skills. The outcomes are enabling CLARIN to prioritise and plan improvements and developments (Sanders 2017).

In addition to teaching how the technologies and services work, an important goal is also to stimulate methodological and paradigm shifts towards integrating qualitative and quantitative methods, interdisciplinary research design, open science policies and transnational collaboration. We hope that the DH community will embrace the CLARIN initiative and will use and extend its resources, tools and expertise.

6 References

Arppe, A., Bruun, S., Koskenniemi, K., Lindén, K., Oksanen, V., and Westerlund, H.: A report including Model Licensing Templates and Authorization and Authentication Scheme. CLARIN-PLUS Deliverable D7S-2.1, CLARIN ERIC, Utrecht, The Netherlands (2011).

Blätte, A. and Blessing, A.: The GermaParl Corpus of Parliamentary Protocols. In: Proceedings of the Eleventh LREC, Miyazaki (2018).

Brugman, H.: CLAVAS: A CLARIN Vocabulary and Alignment Service. In: Odijk, J. and van Hessen, A. (eds.): CLARIN in the Low Countries, pp. 61–69. Ubiquity Press, London (2017).

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. and Schnapp, J.: Digital Humanities. MIT Press, Cambridge, MA (2012).

De Does, J., Depuydt, K., van Dalen-Oskam, K. and Marx, M.: Namescape: Named Entity Recognition from a Literary Perspective. In: Odijk, J. and van Hessen, A. (eds.) CLARIN in the Low Countries, pp. 361–370. Ubiquity Press, London (2017).

De Jong, F.M.G., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, D.: CLARIN: Towards FAIR and Responsible Data Science in the Area of Language. In Proceedings of the Eleventh LREC, Miyazaki (2018).

De Smedt, K.: Some Reflections on Studies in Humanities Computing. Journal of Linguistic and Literary Computing 17, 89–101 (2002).

Edmond, J.: Collaboration and Infrastructure. In: Schreibman, S., Siemens, R. and Unsworth, J. (eds.): A New Companion to Digital Humanities. Wiley, Chichester, UK, pp. 54–65 (2016).

Edwards, C.: The Digital Humanities and Its Users. In: Gold, M.K. (ed.): Debates in the Digital Humanities, pp. 213–232. U of Minnesota Press, Minneapolis (2012).

Fišer, D., Lenardič, J., and Erjavec, T.: Meet CLARIN's Key Resource Families. In Proceedings of the Eleventh LREC, Miyazaki (2018).

Gold, M.K. (ed.): Debates in the Digital Humanities. University of Minnesota Press, Minneapolis (2012).

Goosen, T., Windhouwer, M., Ohren, O., Herold, A., Eckart, T., Ďurčo, M., and Schonefeld, O.: CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. In Selected papers from the CLARIN Annual Conference 2014, October 24–25, Soesterberg, The Netherlands, pp. 36–53. Linköping University Electronic Press (2015).

Hinrichs, M., Zastrow, T., & Hinrichs, E. W.: WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In: Proceedings of the Seventh LREC, Malta, pp. 489–493 (2010).

Kestemont, M., Stronks, E., de Bruin, M. and de Winkel, T.: Did a Poet with Donkey Ears Write the Oldest Anthem in the World? Ideological Implications of the Computational Attribution of the Dutch National Anthem to Petrus Dathenus. In: Abstracts of DH2017, Montreal (https://dh2017.adho.org/abstracts/079/079.pdf) (2017).

Maegaard, B., Van Uytvanck, D. and Krauwer, S.: CLARIN Value Proposition. Public Report CLARIN CE-2017-1093-P001. CLARIN ERIC, Utrecht (2017).

Manovich, L.: The science of culture? Social computing, digital humanities and cultural analytics. Journal of Cultural Analytics (2016).

Martinez-Ortiz, C., Kenter, T., Wevers, M., Huijnen, P., Verheul, J. & Van Eijnatten, J.: Design and implementation of ShiCo – Visualising shifting concepts over time. In: M. During et al. (eds.), HistoInformatics 2016: Proceedings of the 3rd HistoInformatics Workshop on Computational History, 11–19 (2016).

Moretti, F.: Graphs, Maps, Trees: Abstract Models for Literary History. Verso, London and New York (2005).

Moretti, F.: Distant Reading. Verso, London and New York (2013).

Nguyen, D., Doğruöz, A., Rosé, C., and de Jong, F. M. G.: Computational sociolinguistics: A survey. Computational Linguistics 42(3), 537–593 (2016).
Odijk, J.: Introduction to the CLARIN Technical Infrastructure. In: Odijk, J. and van Hessen, A. (eds.): CLARIN in the Low Countries, pp. 33–44. Ubiquity Press, London (2017).
Ravenek, W., van den Heuvel, C. and Gerritsen, G.: The ePistolarium: Origins and Techniques. In: Odijk, J. and van Hessen, A. (eds.) CLARIN in the Low Countries, pp. 317–323. Ubiquity Press, London (2017).
Rheault, L., Beelen, K., Cochrane, C., Hirst, G.: Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. PLoS ONE11(12): e0168843 (2016).
Sanders, W.: Focus Group on User Involvement, conducted during the CLARIN-PLUS Workshop "Working with Parliamentary Records", Sofia, Bulgaria, 27 March 2017. CLARIN report (https://office.clarin.eu/v/CE-2017-1091-Focus-group-UI-2017-03-27.pdf) (2017).
Schreibman, S., Siemens, R. and Unsworth, J. (eds.): A New Companion to Digital Humanities. Wiley, Chichester, UK (2016).
Science Europe: Presenting a Framework for Discipline-specific Research Data Management. Science Europe Guidance Document D/2018/13.324/1 (2018).
Snelders, S., Huijnen, P., Verheul, J., de Rijke, M. and Pieters. T.: A Digital Humanities Approach to the History of Culture and Science: Drugs and Eugenics Revisited in Early 20th-Century Dutch Newspapers, Using Semantic Text Mining. In: Odijk, J. and van Hessen, A. (eds.) CLARIN in the Low Countries, pp. 325– 336. Ubiquity Press, London (2017).
Svensson, P.: Humanities Computing as Digital Humanities. Digital Humanities Quarterly 3(3) (2009).
Terras, M., Nyhan, J. and Vanhoutte, E.: Defining Digital Humanities. A Reader. Ashgate, Farnham (2013).
Tiepmar, J., Eckart, T., Goldhahn, D. and Kuras, C.: Integrating Canonical Text Services into CLARIN's Search Infrastructure. Linguistics and Literature Studies, 5(2), 99–104 (2017).
Trippel, T. and Zinn, C.: DMPTY – A Wizard for Generating Data Management Plans. In: Selected Papers from the CLARIN Annual Conference 2015, Wrocław, Poland, Linköping University Electronic Press, 71–78 (2016).
Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., Beunders, H.: The debates of the European Parliament as Linked Open Data. Semantic Web 8(2), 271–281 (2016).
Van der Aalst, W. M. P., Bichler, M., and Heinzl, A.: Responsible Data Science. Business & Information Systems Engineering, 59(5), 311–313 (2017).
Van Ossenbruggen, J.: Trusting Computation in Digital Humanities Research. ERCIM News 111, 23–24 (2017).
Van Uytvanck, D., Stehouwer, H., and Lampen, L.: Semantic metadata mapping in practice: The Virtual Language Observatory. In: Proceedings of the Eight LREC, Istanbul, Turkey, pp. 1029–1034 (2012).
Verberne, S., van Leeuwen, R., Gerritsen, G. and Boves, L.: RemBench: A Digital Workbench for Rembrandt Research. In: Odijk, J. and van Hessen, A. (eds.) CLARIN in the Low Countries, pp. 337–350. Ubiquity Press, London (2017).
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific data, 3(160018) (2016).
Wynne, M.: The Role of CLARIN in Digital Transformations in the Humanities. International Journal of Humanities and Arts Computing 7(1-2), 89–104 (2013).
Zinn, C.: The CLARIN Language Resource Switchboard. In: Abstracts of the CLARIN Annual Conference 2016, Aix-en-Provence, France (https://www.clarin.eu/sites/default/files/zinn-CLARIN2016_paper_26.pdf) (2016).