The Role of CLARIN in Digital Transformations in the Humanities

Martin Wynne
University of Oxford
martin.wynne@oerc.ox.ac.uk

Abstract

CLARIN is a recently-established research infrastructure which aims to build and sustain services based on language resources and tools. CLARIN aims to support and foster the next generation of research in the humanities, which will make use of advanced digital technologies. A distributed infrastructure is necessary in order to overcome the problems of the current fragmented environment, to create an ecosystem in which data and tools can be connected, and in which innovation will be encouraged. Case studies of early CLARIN demonstrators give a flavour of the possibilities of digital transformations in a number of humanities disciplines, and there is huge potential for important future new directions in literary and linguistic computing. For more widespread, thoroughgoing and effective transformations to take place, builders of infrastructure and researchers will need to negotiate and avoid potential pitfalls, and agree to achieve a certain measure of consensus in terms of priorities, categories and concepts. In the context of current debates about the nature of the humanities and their role in society, it will be necessary for digital humanists to be careful to preserve the unique character and importance of research in the humanities, while moving towards research infrastructures which will facilitate digital scholarship.

Introduction

The recent establishment of the CLARIN European Research Infrastructure Consortium (ERIC) is a landmark event in the organization and support of research activity in Europe. CLARIN is building networks and services with the goal of facilitating the next generation of research in the humanities, transformed by enhanced digital resources and methods.[1] What will this research look like? What new questions are starting to be addressed? What are the potential dangers and pitfalls, and what steps will be necessary to make sure that the potential for transformative research is realised?

CLARIN is the Common Language Resources and Technology Infrastructure, and among the various established and emerging research infrastructures, it is distinguished and defined by its disciplinary scope (humanities and social sciences), datasets, tools and services (those relating to language), and geographical scope (European). The goal of CLARIN is to turn the existing, fragmented technologies and resources for processing and analysing human language into accessible and stable services. The purpose of the infrastructure is to offer persistent services that are secure, and to provide easy access to language processing resources. As language, speech and vision technology improve, it should be commonplace to ask questions such as: 'list all uses of “enthusiasm” and derived forms in English novels of the 1840s written by women', 'find all references to the “state” and “*l'État*” with variant spellings and synomyms in political and philosophical texts before 1532', and 'what were the frequencies and varying characteristics of discourse

about the Holocaust in the Western media in the post-war period?'. But in the absence of access to the necessary infrastructure, the technologies to make these tasks possible will only be available to a few specialists, if at all. At present, these questions are very difficult for the average scholar in most humanities disciplines to address.

Ultimately, CLARIN aims to facilitate the posing and exploration of important questions such as these, but much of the effort and the discussion so far in building CLARIN has concentrated on the various practical barriers to building research infrastructures, such as the legal, administrative, financial and technical issues which have to be addressed. CLARIN has focussed on activities such as setting up new consortia, the technical interoperability of tools and datasets, and federated access and identity management. These activities are essential preconditions for building an effective and sustainable research infrastructure, but are not necessarily engaged with the researchers who are intended to be the beneficiaries, and not directly focussed on the research questions that the new facilities will enable. This apparent oversight has been necessitated by the funding regime, and also by the logical priority of overcoming certain practical hurdles before even simple demonstrators could be built.

Early engagement with Humanities researchers

Despite this early focus on legal, technical and administrative hurdles, user requirements have been investigated in various ways, and users have been engaged in various of the development processes. In the CLARIN Preparatory Phase, in the period 2008-10, the developers of CLARIN worked with numerous projects, in order to engage with the ultimate target research communities. These projects included long-standing research humanities programmes such as the circulation of knowledge in the 17th-century Dutch Republic ('CKCC'), the Dictionary of Danish Insular Dialects, and historical text mining for assisting the study of discourses in Early Modern England. Alongside these, there were projects on linguistic topics ('Hierarchical Lexical Functions relating to Proper Nouns'), new research questions ('Consumption patterns and life-style in Swedish novels 1830-1860'), innovative methods ('Narrative Social Psychological Studies of European History'), and inter-disciplinary studies ('Diagnostis And Terminology In Speech Therapy').

As CLARIN moves into the phase of building the proposed infrastructure services, most of the funding and the activity is taking place at the national level, with important national initiatives to build infrastructure

components, which need to be coordinated and integrated to provide coherent Europe-wide services. The CLARIN ERIC coordinates this integration effort. Engagement with research communities at this relatively early stage in construction continues to take place mainly for the purposes of eliciting more detailed user requirements. For example, CLARIN-D, the German national initiative, has discipline-specific working groups covering philology, linguistic fieldwork, anthropology, language typology, ancient history, classical philology, archaeology, human speech processing, psycholinguistics, cognitive psychology, speech sciences, and applied and computational linguistics. The aim is two-fold: firstly, to ensure the conversion, transition, and, if necessary, creation of the relevant tools and datasets to allow these disciplines to participate in CLARIN; and secondly, to ensure the input of researchers who will be among the ultimate end-users of CLARIN services into the development of those services, to ensure that the next generation of research to which they aspire will be enabled. CLARIN-NL in the Netherlands has a set of well-developed demonstrator projects with working practices which are already being embedded in the emerging infrastructure. These two national initiatives are drawn on for the case studies below.

In order to realize more fully this potential for new research projects, CLARIN has identified numerous hurdles which need to be overcome. These include more numerous and improved resources, tools and services, better handling of copyright issues, and more standardization of resources. Much can be said about each of these issues, but the key insight of CLARIN has been to identify the landscape in which it will be possible to address them most effectively, and this will be examined in the following section.

Distributed infrastructure to support research in the Humanities

The CLARIN vision is for a distributed research infrastructure, where there is no single centre, but rather a suite of services hosted in different centres, which can work together to produce a whole which is more than the sum of the parts. The assumption is that such an architecture is the most appropriate for achieving the high levels of interoperability and sustainability necessary to enable digital scholarship with language resources. The emerging paradigms of research computing, characterized by fast and high capacity networks, a deluge of data, and services in the cloud mean that it is increasingly possible to imagine and build distributed architectures for scholarly services, where data, tools, computer processing and the outputs of annotation and analysis live in different parts of the network but can be brought together virtually in the

user's desktop environment.

There has been for some years a high degree of concern about the creation of 'digital silos', in which the outputs of digital humanities projects are deployed online in such ways that they are unconnected to other resources, and with limited sustainability. This problem is directly addressed by the vision of a distributed infrastructure.

The CLARIN vision is that one day a researcher, located anywhere in the world, from his desktop computer, will be able to:

- perform a single sign-on, with local authentication, and then:
- search for, find and obtain authorization to use corpora in located in a number of different repositories;
- select the precise dataset to work on, and save that selection;
- run semantic analysis tools hosted in one centre and statistical tools from another in a newly defined workflow;
- use computational power from the local, national or other computing centre where necessary;
- obtain advice and support for carrying out all technical and methodological procedures;
- save the workflow and results of the analysis, and share those results with collaborators wherever they have affiliations and are physically situated;
- discuss and iteratively adopt and re-run the analyses with collaborators.

An open and fully distributed architecture where the resources are located in different places is necessary to this vision. Such a form of organization can have the advantages of allowing the services to be created which are not limited by the resources, imagination and expertise of an individual resource provider. Building *ad hoc* collections and corpora across different repositories becomes possible, as do complex workflows which might pipe together web services from different locations. Protected resources, whose

distribution and access might be restricted due to copyright, or other legal or ethic factors, can be curated and protected *in situ*, yet still analysed online by granting limited and controlled access via web applications which access the data in a secure environment.

All of this can happen in a situation with a more efficient division of labour than typically exists today. Rather than the vertical integration which means that each centre has to provide specialist services for data curation and use, in this scenario, the repositories do not have to worry about tools; tool and content developers don't have to worry about creating the entire online environments; tool developers don't have to worry about data management; users don't have to install software. The emergence of an 'ecosystem' with numerous actors providing content, tools, computing resources, and other infrastructure services, provides a flexibility and resilience and the potential for sustainability which is not possible for a single site or other more closed or monolithic systems.

The services which are available now via CLARIN include WebLicht for resource processing and workflow management; the Virtual Language Observatory for resource discovery; tools to support resource creation and enhancement; European Persistent Identifier Consortium (EPIC) service; repository services in the centres for archiving, preservation and sharing; federated identity management (including a CLARIN Identity Provider and a service provider federation with international agreements in place). Services which will be available in the future, all of which are currently under development and at least in the alpha release stage, include federated content search, monitoring of online availability, a registration and certification system for centres, a virtual collection registry, virtual workspaces and cloud storage, and a safe replication service.

This is no simple panacea, and numerous difficulties emerge, many of which relate to access and identity management. Arranging access to services in one location can be hard enough, but authorization to use textual data in more that one repository requires the passing of identity information between multiple institutions. While there are reasonably well-established technologies and procedures and agreements for controlling access to online content, the authorization of web services is not such a well-established area. Furthermore, authorization to access online content cannot easily be passed on to authorize access to the computer processing power that is necessary to carry out the analysis of the content, especially if this is

being provided by another centre in the distributed infrastructure. In summary, the fact that distributed services are reliant on cross-institutional agreements and arrangements adds an extra hurdle to be cleared in order to make the system work. The assumption behind CLARIN's approach to these problems is that they need to be overcome by high-level agreements between national federations, because they cannot be addressed at the level of the individual researcher, or project, or service provider, or even institution. Agreements to establish domains of trust between institutions can allow researchers to use their local online identity to access resources online, and are necessary for such scenarios to function effectively.

In summary, the CLARIN infrastructure aims to provide a range of simplified solutions for connecting together data and tools, deploying them as reliable services, managing authentication and authorization, gaining access to computing power, monitoring availability, and making connections to virtual research environments. The real effectiveness of the infrastructure will depend on its take-up by researchers and its population with interoperable tools and resources. This will not be easy, and will require the investment of time by researchers and developers to learn the necessary protocols and habits, and a certain re-orientation of working practices, including a willingness to accept certain compromises in terms of adoption of standards or idealization of data models.

The distributed technical architecture of CLARIN will be an infrastructure to support researchers throughout the life-cycle of their work, which has to be situated in the context of a broader picture of infrastructures, in which CLARIN is only one part of the jigsaw. While CLARIN focusses on language resources, technologies and services, DARIAH will also provide complementary services to humanities disciplines and communities, and DASISH explores the numerous cross-community services which can be built. Parts of the infrastructure are already in place, with compute Services provided by DEISA and the EGI, data services from EUDAT and network services from GEANT[2].

CLARIN is building a part of the research infrastructure with services that are flexible but standards-based, and with a distributed, federated architecture, integrated into a wider ecosystem of infrstructure services. There are many technical challenges, but if they are to be addressed in ways which will respond to the requirement of users, then it is not too soon to ask what new forms of research might be made possible. The next section will examine a number of CLARIN demonstrator projects, and reflect on the type of research

questions which are being addressed.

Case studies of Humanities research

War in Parliament (WIP)

War in Parliament (WIP) is a demonstrator project which is making available the digitized collection of the proceedings of the Dutch parliament (Handelingen der Staten-Generaal) from the period 1930-1995 as a fully annotated dataset, and an advanced search engine in order to make it possible to carry out historical and social science research. The data and tools are being prepared in such a way that they follow the emerging standards, formats and protocols to make them usable within the CLARIN infrastructure. The project team are focussing on research questions relating to references to World War II, to examine and explore how such references were deployed in parliamentary discourse.

The Second World War has been a powerful benchmark for political morality in post-war Western societies. Political elites have felt a constant need to relate to the defining experiences of war, defeat and genocide. War in Parliament is a study of the impact of World War II in post-war political debates and decision-making in the Netherlands. The assumption is that an analysis of parliamentary debates will reveal how the political elites constructed narratives evoking the wartime past of their nation to make sense of the present. A case study of the debates surrounding the *Boerenpartij* (Famers Party) will shed light on the way references to World War II were used within political discourse, and at the same time demonstrate the power that digital research can bring to historical and social science research.

The *Boerenpartij* was the first political party from the far right to enter Dutch parliament after World War II, and was stigmatized by the traditional political parties, with attempts to cast it outside of the acceptable norms of politics, and associate it with National Socialism. The contention of the researchers is that no systematic research has been conducted on the question of how frequently, in what ways, and for what purpose, the discourse of parliament explicitly linked the *Boerenpartij* with Fascism and National Socialism, and to what extent the Nazi past was used as a political weapon. The researchers also plan to broaden their research questions to answer questions about how discourses have evolved over a longer period of time. Dutch historiography on the aftermath of World War II shows that the persecution of Jews comes

increasingly to the fore in national discourse towards the end of the 1960s which challenges a more pervasive popular idea that the Netherlands was seen as a country characterized by resistance to the German occupation throughout the wartime and post-war periods.[3] With the search engine developed in this project and a comprehensive digital dataset of parliamentary debates, it becomes possible to unearth forgotten and unexamined debates, thus contributing to new insights into the legacies of the Second World War.

The War in Parliament project has made extensive use of XML technologies to transform the scanned images of the Dutch Hansard into a semi-structured dataset. The texts have been annotated with hierarchies of topics, scenes, and speeches. Databases of Dutch politicians and political parties were created from publicly available sources and used to annotate the speeches from 1930 to 1995 with a speaker name, role, and party affiliation. All datasets are made available in well-formed and validated XML documents, and can be queried with the use of XPath expressions or XQuery. A search engine was developed for simple text searches, with the ability to restrict output by speaker name, party, date range, and other criteria. The use of standard technologies for encoding and search, along with the embedding of the data and tools in the CLARIN infrastructure, should facilitate the reuse of data and tools by a wide range of scholars via services which will remain sustainable in the long term.

As well as examining these questions, the results which are being derived from War in Parliament invite scholars to think about the differences between traditional history and historical research in the new digital era. To what extent will these new methodologies transform historical research? Will digital approaches contribute to improvements in the confirmation and reproducibility of historical scholarship? It will also be interesting to see whether the availability of reliable tools makes it easier for non-specialists and those outside of the academy to use these resources, which are of interest to journalists, writers, politicians and the general public, as well as academic historians.

WAHSP: Towards a flexible and stable CLARIN-supported webapplication for historical sentiment mining in public media

This project aims to make use of the CLARIN infrastructure and to populate it with large historical datasets of newspapers and journals, plus the necessary tools to investigate opinions about the use and abuse of drugs between 1900 and 1945. The challenge is to convert a specific text mining technology ('sentiment mining')

into an accessible, CLARIN-compliant web application which can be used to address the research questions brought forward by historians and social policy researchers. The interdisciplinary project team of historians, linguists, and computer scientists are converting existing tools to the specific needs of digital humanities research. The development of this demonstrator will also be used to refine a list of requirements to which the CLARIN infrastructure should respond.

WAHSP aims help historians to learn more about public sentiments in respect of drug use in this period, exploring a wider public beyond the well-documented discussions, opinions and actions of policy makers and law enforcers. WAHSP focuses on public opinion as expressed in the news media in the Netherlands, and also in its overseas colonies. In this way, WAHSP aims to explore and make visible the hitherto hidden references to the use and abuse of drugs in the wider discourses of society, not just in the articles or commentaries which are explicitly on the topic of drugs. The intention is that sentiments which are widely held, yet not necessarily under discussion, can be revealed.

The search engine searches through the digital newspaper collection of the Netherlands Royal Library from the period between 1619 and 1995. The collaboration between the scholars from several disciplines has resulted in a semi-automatic and interactive open-source query-based search engine to extract the relevant data. It is only semi-automatic because the efficacy of the information extraction depends on the expertise of the scholar to come up with valuable keywords. The application analyses the results in a number of ways. Sentiments and evaluative language can be highlighted. A quantitative analysis of the sources used can give insights into social contexts if their use. Word clouds provide visualizations of the contexts in which certain keywords were used in articles, and can be generated from the results of querying hundreds of articles together.

WAHSP is an important example of the work of CLARIN. It is about converting existing datasets and tools so that they can be combined and interoperate and be used by a group of scholars. This is an example of CLARIN working to remove the barriers to addressing existing research questions with digital resources.

VisArgue – Why and when do arguments win?

A team of linguists, information scientists and political scientists in Germany is investigating the question

of when political negotiations are successful, and why. In order to carry this out, they are developing tools for the automatic analysis of political discourse, allowing the human analyst to make interpretations about the effectiveness of the discourses of discussion, consultation, argumentation and decision-making.

The team is motivated by the observation that large-scale publicly funded construction and infrastructure projects in Germany in the recent period have created conflicts between governments and civil society, and the realization of these projects has become a huge risk for political decision makers. The researchers aim to investigate the factors that make political communication successful. This complex task can only be tackled using an innovative combination of methods from different disciplines. These methods include a deep and detailed linguistic processing of verbal mediation processes in order to generate an abstract representation of communication; a shallow, statistical analysis of text to detect common patterns in negotiations; and the development and employment of visualization tools which identify patterns of communication for human analysis.

As well as shedding light on a topical social question, the analysis of these debates allows for an empirical testing of the specific theory of deliberation which the researchers are applying. This thorough, verifiable and reproducible testing of the theory would be difficult without the large-scale, formal analysis of digital texts. The main objective is to develop an automated framework that measures the quality and effectiveness of deliberation. This framework is based on the detailed linguistic processing of discourse relations and informational structures, which involves mapping political discourse onto an abstract level of the representation of communication, based on lexical-functional grammar. This representation is then visualized with statistical approaches to detect common patterns of argumentation in political negotiations. As such, the tools and methods developed in this project, if successful, will be deployed as services in the CLARIN infrastructure for further re-use, development and refinement.

Arthurian Fiction in Medieval Europe

CLARIN is also supporting historical research from the more distant past. *Arthurian Fiction in Medieval Europe: Narratives and Manuscripts* is a research resource which provides information on medieval Arthurian narratives and the manuscripts in which they have been transmitted throughout Europe. The database consists of linked records relating to more than two hundred texts, more than a thousand

manuscripts and two hundred persons. With the publication of this resource, the compilers hope to foster further research into Arthurian fiction as a pan-European phenomenon. The database is work in progress: a considerable number of records have yet to be completed, while fresh discoveries of narratives and manuscripts invite new entries.

The entire corpus of surviving medieval stories about King Arthur and his associates is being studied as a multi-lingual and supranational phenomenon, shedding light on the transmission of narratives in Europe in the period between 1150 and 1550 AD. Arthurian stories constitute one of the most influential genres of western literature, and are studied intensively. It is considered that their fictional character made them ideal vehicles for expressing the social concerns and cultural values of their audiences.[4]

Before the collaboration with CLARIN, the datasets were not available for use online. The datasets were converted to an XML standard, from which metadata is extracted and made available to CLARIN harvesters. The demonstrator allows users access to the data for searching as well as for adding and correcting records. As a result, the Arthurian Fiction web database enables scholars to work online to answer questions on a pan-European level for the first time.

Rethinking Literary and Linguistic Computing in the Era of the Data Deluge

Current CLARIN demonstrators such as those introduced above tend to deal with specific and finite datasets. Such studies comprise the backbone and the bulk of in the humanities today, and CLARIN can help to make them easier to carry out, lowering the hurdles faced by researchers who want to get involved in advanced digital scholarship, by enabling new instances of data linking and new forms of collaboration. However, these examples do not exhaust the full range of possibilities for exploiting the data deluge. The CLARIN infrastructure has not yet been populated with the data and tools to make this possible, and in certain important cases the underpinning services are not yet ready. Supra-national federated identity management systems are necessary to allow secure access to researchers from different countries, but the agreements are not all in place and some of the technologies not yet mature. Publishers of large collections of digitized materials do not make them available via the CLARIN infrastructure. The tools to cross-search disparate and heterogenous collections are not easily available to researchers. Beyond the realm of digital text, the tools and services to exploit other media are underdeveloped. The next generation of digital

transformations is yet to be made, and can only be imagined, but this is not an impossible exercise, and will be briefly attempted here.

The traditions of literary and linguistics computing are largely based on methods of working with shared text collections and corpora which are well-established, well-documented and have been repeatedly re-examined, annotated and analysed. Born-digital data, along with large-scale digitization of historical texts, are delivering the cultural products of the both the present and the past directly to the desktop. New bespoke datasets can relatively easily be swiftly woven for different research questions. The boundaries between the corpus and other type of data are becoming blurred. So the question needs to be posed whether we can justify spending our time carefully crafting and annotating corpora today. This key question poses a further one for CLARIN - in order to bring about the transformations in digital research to which we aspire, are the humanities best served by an infrastructure based on these bounded, finite, discrete corpora, or a more expansive set of tools for mining the wider digital environment?

Nowadays digital text is everywhere, searchable, downloadable, and analysable on the fly. The literary and linguistic disciplines are experiencing the data deluge described for the hard sciences a decade ago, when new instruments and networks started to flood researchers with digital outputs from new instruments and experimental facilities. The sciences have been transformed by this experience. Are we seeing the data deluge transform the literary and linguistic disciplines of the humanities?

It is possible to still make the case for the corpus which represents types of language with which we are not being deluged on the internet - speech, other non-computer-mediated data, historical documents of various types which are not easily available online, and for which they can be confident about their integrity, levels of quality assurance in digitization, and provenance. Many researchers still make the case for the annotated corpus, with added annotations embodying interpretative information.

This all raises questions for researchers in terms of how best to allocate their time and resources. Time spent on carefully constructing, annotating and documenting corpora increasingly risks confining our horizons to ageing research methods, always behind the cutting edge due to the time delays in construction and annotation, and the necessarily finite size of the final dataset. Many researchers see the need to go beyond the finite text corpus in our next generation of research, and to examine huge datasets, as well as

digital media, and the languages and genres of the internet. We also have the possibility of capturing more of the context of language events, and aligning text and speech with other data streams, such as hand and eye trackers, geographical position sensors and even heart-rate and brain activity. Much new research is already exploring the possibilities of working with non-finite corpora, beyond the limits of the repositories in our community, and engaging with the idea of "the web as corpus".

These discussions raise questions of priorities: should CLARIN be populating a research infrastructure by converting existing datasets and tools (which was the original goal), or new tools for mining the web? Does the focus on repurposing existing resources as sustainable and interoperable services inherently lead to a certain conservatism and inhibit innovation? Should we be supporting the detailed annotation of text corpora, when we could be exploring new data types, methods, tools?

The most cogent contributions to these issues argue that we need instruments that will allow both detailed analysis of well-understood data of known provenance, as well as and alongside the more expansive mining of large datasets. The options and techniques here are sometimes characterised as distant, close and scalable reading. There are those who resist the digital humanities and yearn only for the traditions of close reading.[5] On the opposite extreme are those who argue for and pursue only distant reading, among whom might also be counted many corpus and computational linguists whose research methods focus on categorizing, counting and annotating, and not on interpretation of the meaning of texts.[6] Greg Crane, and others, have asked what can we do with a million books.[7] The problem is how to reconcile the deep knowledge of texts, and close reading of them, with broad brush overviews, statistics, digests and trends. A more balanced position, aiming to take advantage of new possibilities, but without discarding the hermeneutic tradition, is what Martin Müller calls 'scalable reading'.[8] New instruments are necessary for scalable reading, and to achieve them would require not only making investments in infrastructure, and making serious efforts towards the interoperability of the resources which the infrastructure makes available.

Annotation is of particular importance to this question, since the addition of interpretative mark-up to texts and other media is one of the principle operations carried out in the creation of digital resources and in their analysis, particularly in the literary and linguistic disciplines. There is the danger that adding annotation is too time-consuming and costly, and that it forces researchers into using smaller and older datasets, and only

starting the analytical phase of research after the corpus has been painstakingly annotated. Furthermore, there is a lack of generic software which can make use of annotations, and so the trend of building separate web interfaces for each annotated corpus continues and is reinforced. Geoffrey Leech has argued that annotations should be separable from the text (i.e. that they can be removed so that the text can be used without them), that detailed and explicit documentation should be provided, that annotation practices should be linguistically consensual, and that annotation should observe standards .[9]

The questions that this raises about consensus and standards are extremely relevant to the priorities of research infrastructures. Without accepted standards, annotation can lead our resources into digital silos, because of the lack of generic tools for exploiting annotations in meaningful ways. But standards in annotation involves more than discussions about coding schemes and tagsets. It requires a certain level of consensus on concepts and categories, even though research in the humanities might be precisely about a critical approach to these concepts and categories. And the humanities are divided into many different communities, with different baseline assumptions and working practices.

The new forms of digital scholarship have emerged in the natural and physical sciences, and speculation about of the possibilities of e-research in the humanities often makes reference to them. It is important to note that these new forms are often underpinned by standards, consensus and compromise. Scientists have come together to agree on some basic categories and ways of representing them, and on ways of expressing reproducible workflows. These compromises and hard decisions have allowed them to find out more, to better approximate the truth, and to accumulate knowledge. New paradigms of data-driven web science are emerging. Instead of putting a paper thesis on the library shelf, or a PDF in the e-prints archive, the output of scientific research could become an executable thesis capable of producing new research, thanks to the use of linked open data and code, and reproducible workflows. [10] But to achieve such new forms of digital scholarship requires a high level of agreement about not just the syntax of data models and annotations, but also about the semantics.

It should be acknowledged at this point that the humanities are not necessarily about quantifying phenomena, building models, accumulating data, and the falsification and reproducibility of research. Perhaps linguistics, as the 'science of language' has led the way towards a scientific approach to the

humanities, and has blurred important distinctions between the methods and the objects of study of the humanities and the sciences by following an excessively and inappropriately scientific approach to the humanities. The conundrum then for the digital humanities is that it is difficult to achieve transformations in digital research while we remain committed to close reading and the hermeneutic traditions, and while we continue to allow the constant questioning of our fundamental assumptions and categories. If we want to realize the vision of a digital humanities which can address new, and more expansive questions, and which can do distant reading as well as close reading, it will be necessary to make some compromises - which can be provisional, *ad hoc*, subject to constant examination and discussion - in order to move towards more semantic operability of our resources.

Conclusion: Digital Transformations and the value of the Humanities

The questions of strategy and priorities for building infrastructure and transforming research can be considered in the context of current discussions about the humanities in society. There are currently a number of heated debates about the nature of the digital humanities, the role of the humanities in education and in society more generally, and, indeed, about the current relevance and status of the Enlightenment project to understand humanity and society.[11] The humanities are being asked to justify their existence in instrumental terms by funders and politicians. A dangerously seductive route lies open to those involved in CLARIN, which could allow them to argue for the instrumental, vocational and economic value of language technologies, and to point to how advances in corpus linguistics have contributed to, for example, better and more profitable learner dictionaries and a new economy of small, start-up, web-based language technology firms. To argue for a digital humanities which is primarily concerned with the accumulation and analysis of data, and which has goals of promoting the economy, and other specific social or political goals (such as fighting terrorism) would do a disservice to the humanities. As Stanley Fish and others have noted, there is a priority of the humanities to stand up for its own traditional values today.[12] Society needs the humanities to help us understand the things that science can't tell us. In the era of the data deluge and web science, we need to constantly question our working methods.

> [There is] a monolithic conception of social space, according to which it would suffice to have the right information to make the right decisions. But in point of fact, information itself is far from

> homogenous and no purely quantitative approach is satisfying. Having ever greater amounts of information at our fingertips not only does not make us more virtuous, as Rousseau already predicted, but it does not even make us more knowledgeable.[13]

For Todorov, studying humanities can teach us about virtue as well as truth. Science and the humanities can tell us different things. Science can seek the truth about phenomena., while the humanities can tell us what is good and what is valuable. Is it the case that the digital humanities only push us towards the search for scientific truth, with the danger of "scientism", an excessive reliance on science, to try to answer questions about morals, ethics, how to live our lives and how to be good? CLARIN must avoid leading the way to such a "scientistic" approach to the humanities, reducing questions of values and morality to counting features, and seeking instrumental or profit-oriented impact and results.

The challenge for CLARIN is to create an ecosystem in which better hermeneutically-informed software can be created and deployed. This would enable the sharing of resources with provisional, *ad hoc,* but agreed categories for representing our analyses and interpretations. Criticism and scholarship relating on the nature of the interpretations implicit in these annotations would not come to a halt, but achieving agreements such as these would remove barriers to the creation of large-scale shared digital facilities.

We need to follow the sciences in deciding priorities, adopting standards, reducing complexity and variety, but only as pragmatic measures to promote shared facilities and infrastructures. At the same time, we need to avoid the promotion of an excessively data-driven, empirical and scientistic view of the humanities, and continue to defend the traditions of qualitative research in the humanities, and pursue the humanities for their own sake.

Acknowledgements

The CLARIN Preparatory Phase project received funding from the European Community's Seventh Framework Programme. CLARIN-D receives funding from the Federal Ministry for Education and Research in Germany, and CLARIN-NL from the Ministry of Education, Culture and Science in the Netherlands.

Endnotes

1 T. Váradi, S. Krauwer, P. Wittenburg, M. Wynne and K. Koskenniemi, 'CLARIN: Common Language Resources and Technology Infrastructure', Proceedings of the Sixth International Conference on Language Resources and Evaluation (Malta, 2008).

2 See http://www.clarin.eu/ , http://www.dariah.eu/, http://dasish.eu/ , http://www.deisa.eu/ , http://www.eudat.eu/ , http://www.egi.eu/ , http://www.geant.net/ , all last accessed 20 February 2013.

3 Not just in the Netherlands, according to the controversial thesis of N. Finkelstein, *The Holocaust Industry* (London, 2000).

4 K. Busby, B. Besamusca and F. Brandsma, eds., *Arthurian Literature: The European Dimensions of Arthurian Literature* (Woodbridge, 2007).

5 S. Fish, 'Mind Your P's and B's: The Digital Humanities and Interpretation', 23 January 2012, http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/, last last accessed 20 February 2013.

6 F. Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London, 2005).

7 G. Crane, 'What Do You Do with a Million Books?', D-Lib 12:3 (March 2006), http://www.dlib.org/dlib/march06/crane/03crane.html, last accessed 20 February 2013.

8 Scalable Reading, https://scalablereading.northwestern.edu/, last accessed 20 February 2013.

9 G. Leech, 'Adding Linguistic Annotation' in M. Wynne, ed., *Developing Linguistic Corpora: a Guide to Good Practice* (Oxford, 2005), 17-29.

10 D. de Roure, S. Bechhofer, C. Goble and D. Newman, 'Scientific Social Objects: The Social Objects and Multidimensional Network of the myExperiment Website', in *1st International Workshop on Social Object Networks,* (Boston, 2011). http://eprints.soton.ac.uk/id/eprint/272747, last accessed last accessed 20 February 2013.

11 See M. Gold, ed., *Debates in the Digital Humanities (*Minneapolis and London, 2012), J. Bate, ed., The Public Value of the Humanities (London, 2011), M. Nussbaum, Not for Profit: Why Democracy Needs the Humanities (Princeton, 2010) and T. Todorov, *In Defence of the Enlightenment* (Kindle edition, 2010).

12 S. Fish, *Save the World On Your Own Time,* (New York, 2008).

13 Todorov, *In Defence of the Enlightenment*.