CLARIAH: BIG DATA, GRAND CHALLENGES

Researchers into the Humanities receive a 12 million grant from NWO (the Netherlands Organisation for Scientific Research) to build a digital infrastructure. This enables them to interpret and disclose large data files consisting of texts, audiovisual material and archive material. But which are the research questions that can now be addressed?

Humanities researchers have long studied human culture. They ask themselves fundamental questions such as: why have certain regions in the world been rich for so long and others poor? Why is it that in the public debate the representation of certain minorities is so rigid? How does language change as a result of migration? For decades these questions have been put and answered by historians, media experts, linguists and many other researchers. Scientists with a background in the humanities are very good at interpreting *content*, and especially *distinct data* each *in his own field*. Historians work with structured data from archives. Media experts use text, sound and images from newspapers, television and other (social) media as their source material. And linguists draw from vast textual and oral data files.

One might say that humanities researchers study the *building stones* of culture and *patterns* of cultural change in their own individual way. The building stones with which they have been working for ages (text, images, sound and historical data) are numerous and scattered. That is why many often focus on one piece of the jigsaw to interpret and analyse that as well as they can. For instance, the work of one particular painter, the novels by one particular author, the numerical data from urban archives of one historical period or the language used by one social group. So far, humanities scholars have been less successful in combining these many questions and studying them in an integrated fashion. Examples of such complex questions are: How do differences between European cultures result in processes of in- and exclusion? And how can we learn from history to put modern society on the right track? Such daunting issues require and cohesive understanding of changes in language, socio-economic status, and representation of specific groups. Not as separate phenomena, but as a complex whole.

Big Data

This big challenge can now be taken head on. For just under a decade our heritage centres (archives, libraries, centres of knowledge) have had a growing number of data files at their disposal. The volume has increased in such a way that we have started to use the term Big Data. But in order to mine this wealth of material, new tools will have to be developed: tools to query these data for meaningful content. As a result humanities researchers will not only change their working methods, but they will also be able to pose, and hopefully answer, questions in new ways. To mine these data and to see the connections between the interpretations, researchers have taken the initiative to develop a common infrastructure. The jigsaw pieces in the field of language, (moving) images, sound and historical data need to be put together; experts will have to learn from each other how they can include these data files in their research.

CLARIAH (Common Lab Research Infrastructure for the Arts and Humanities) is a common project of a core team of scientists, supported by a consortium of 40 knowledge and heritage centres, public organisations and companies rewarded by NWO with a generous 12 million grant. This money for a common infrastructure will not only be helpful for the development of digital tools for mining large data files; by making these tools "talk" to one another, humanities researchers will learn to co-operate to answer these complex questions. Three disciplines form the vanguard of CLARIAH: linguistics, media studies and socio-economic history. Linguists focus on mining digital text files. Experts in Media Studies mainly develop tools for interpreting audio-visual sources (sound and images). And socio-economic historians concentrate on structured data files from archives. However, the tools to be developed in various disciplines need to be useful for all researchers working with various types of digital data. A recent survey in de Groene Amsterdammer among 200 humanity scholars in the Netherlands showed that they thought *Digital Humanities*, another name for this watershed, the most important development in their discipline.

Apart from being a boost to *Digital Humanities*, CLARIAH also aims at making a substantial contribution to important scientific questions with reference to Big Data *outside* the humanities. This project will yield building stones that are complementary to the work of exact and social disciplines in the field of data mining. While information scientists are experts at designing search algorithms, and social scientists are curious about user behaviour, humanities are good at *interpreting human messages*. Big Data in the humanities are mainly “rich data”: they are full of noise, in the same way that culture can be fuzzy. Statistics on poverty are no facts, but need interpretation. Opinions in public debate are many but also diffuse – they have a different specific gravity. And images and text can be ironic or ambiguous. Whoever studies culture knows that content needs interpretation and that messages are only meaningful when interpreted in context. Understanding this *complexity of content* – that is what humanities researchers contribute to the study of large quantities of digital data. CLARIAH stands for an even more intensive co-operation between the humanities, exact sciences (especially information technology) and social sciences in understanding cultural complexity.

Grand challenges

But it is not only scientists that stand to gain by the future infrastructure. Industry is also interested in projects that can be evolved with the help of CLARIAH. Perhaps companies do not have a direct interest in solving Big Questions that are posed by humanities researchers, such as the study of migration and minorities. However, they do show an interest in the *type* of knowledge that is being developed here: interpreting complex messages in digital environments. In anticipation of this infrastructure researchers have co-operated with medium and small enterprises in the field of language and speech technology and image recognition. A big company like IBM has shown interest in the contribution of humanities researchers. Talks between IBM, the Royal Academy of Arts and Sciences (KNAW) and the Amsterdam universities about co-operation in Digital Humanities are at an advanced stage. And these forms of public-private co-operation tie in nicely with initiatives in the Top Sector Creative Industry—an NWO funded program in which the Humanities have done very well in the first round

of applications. In short: this infrastructure is an enormous impulse to the co-operation of scientists and industry.

The fact that Dutch humanities researchers have supported CLARIAH so unanimously is fairly unique. Over the past five years researchers from different disciplines – linguistics, socio-economic history and media studies in particular – have made significant progress in their own field, for instance with projects such as CLIO-Infra, HSN, NederLab and EU Screen. Through these investments the Netherlands have gained an extremely strong position in Europe. CLARIAH does not only interconnect Dutch subprojects, but also guarantees a co-ordinated contribution to the more far-reaching European infrastructures DARIAH and CLARIN – infrastructures in which the Netherlands, supported by the Ministry of Education and NWO, plays a major role.

In the short run the emphasis is on the development of tools and the training of researchers in the world of large data files, which is new to many of them. The insight and skills thus gained will enable researchers to address major questions about culture and cultural change. The question what constitutes European identity is no longer "too big" to be studied. Rather, the wealth of digital data and their integrated interpretation will yield answers to this question. The example of inclusion and exclusion in Europe mentioned before has not been chosen arbitrarily. One of Grand Challenges formulated by the European Commission for the scientific agenda *2020* is "Europe in a changing world: Inclusive, innovative, and reflective societies." With the support of the CLARIAH infrastructure Dutch humanities researchers are ready to take up this challenge together, and that is the first time in history.

Grand and small ambitions

CLARIAH itself is not the answer to the major question concerning complex human culture, in the same way that Big Data is not the answer to all kinds of challenges in society. CLARIAH is an indispensable infrastructure for the humanities. In the same way that the Hubble telescope is necessary to solve the riddles of the cosmos, and the MRI scanner to unravel the human fabric, CLARIAH will contribute to investigating the complex meaning of human culture. And while humanities researchers are trying to attain these ambitious goals, they cherish a more down-to-earth hope that their search for an interrelation between text, sound, image and other data will yield many useful tools.

http://www.clariah.nl

Prof. dr. José van Dijck
Professor of Comparative Mediastudies
University of Amsterdam
Department of Mediastudies
Turfdraagsterpad 9
1012 VT Amsterdam
The Netherlands
Phone: 31 20 5252980
E-mail: j.van.dijck@uva.nl
http://home.medewerker.uva.nl/j.f.t.m.vandijck/