DARIAH Winter School in Prague

Open Data Citation for Social Sciences and Humanities

24th to 28 of October 2016

Session 10: Social impact

10-Social impact

Gregory Crane is an Alexander von Humboldt Professor of Digital Humanities at Leipzig University.

In the 1980's, we thought that Greek and Latin texts needed to be online, in an open way, so people could use them with the goal, as being humanist, to transcend transnational cosmopolitanism. Since then, interesting projects have emerged and improved our knowledge. There is for example the coordination project OCR-D, which is aimed at the development of methods of Optical Character Recognition (OCR) for printed historical material. We can today try to extend these tools to quotation detection, networks and mapping, or the analysis of social network data.

Open Greek and Latin, a subset of a global philology project

Greek and Latin remain the central topic of the great national trends of the cultural heritage of Europe. I would like to emphasize the importance of Latin as a transnational aspect of European identity, as an ideal of unified language, which all Europeans can share and which is, on the same time, nobody's language. Scholars and scientists can help transcend short-term political issues. The rise of the great national vernacular languages is very interesting as it shows the choice between using a national language and having a limited audience or the adoption of a language with a broader impact. Open Greek and Latin is in a sense a large-scale open data project as we have reached the first million OCRed books downloaded from Internetarchive.org and this allows to trace translations over time.

Philology

Philology is the cognitio, the mental understanding process of the universal past, historical and philosophical. But it can also be understood in a narrower way as being the preparation of editions, the analyze variances, etc. I see it very differently: anything you can do with a textual record that allow to reconstruct anything that happened in human mind or in the world around us. By definition, it is expensive, never satisfied and has no inherent methods. Methods can evolve depending upon the question. And statistical methods can help our understanding of the past. So, philology is also data driven, every statement is directly backed with the primary evidence which upon the statement is based. Within philology, mechanism for citation is essential as you have to be able to reference any word, symbol, element of any surviving text or object (see Homer Multitext Project documentation)

Humanities and visualisation

1. Difference in the proof and discovery in the Humanities opposed to other scientific analytic fields
2. Evaluation for humanities questions that may have no ground truth?
3. DH and text visualisation scenario

We today have a lot of interesting emerging projects and tools that can address text analysis for data sets or social media, in a multilingual way, but the question is about the usage of this tremendous potential. What kind of questions can you ask given this level of heterogeneous and vast data?
You now have huge databases, such as Internet Archive.org, Europeana, Gallica from the French National Library, etc., but the question then is the connection with national databases on a larger scale. Germany has digitised and carefully scanned about 400 000 books printed in Germany or in German outside of Germany from 1500 to 1800. When it is OCRed, the implications would transform the way we have to think about the history of thought, as we would have a bigger scale. We now have tools that allow the visualization of automatically detected text reuse for a document in millions of books, which is absolutely nice for the humanities. This is really something that should happen to all our documents.
=> Digital Humanities are the space of creative destruction where students of the humanities are forced to rethink their larger goals in light of the challenges and possibilities of the digital world, beyond the fixed stream of print on the page.

One opportunity with big data is the use of geotags and place-names as it allows to generate a map from the content of a document, and its relations, it allows to see things you couldn't see before.
=> This is why it is interesting to get access to the source data, but, the problem with big data is the multiplicity of languages and the volume, which cannot be done by hand anymore.

You can now produce visualisations of cluster of words that statistically co-occur. These clusters often correspond to a topic, but the ground problem then is to know what you can do with that. You can also use bi-lingual text display tool that helps understand the structure of a text with grammatical and syntactic relation of every word; you can soon figure out how to read a text in another language. This is an environment where you can do something with a text that 30 years ago was completely inaccessible. This is beyond translation because you can see all the functions of words. This is one way to understand data and open the barrier of language for inspection. And it can also be applied to music as a textual object (with score), or to mathematics.
How far can you get? How do you generate data that you need to understand? How do you do syntactic analysis at scale? How do you make parallel text alignment automatically?
We have examples of students and citizens who are voluntarily producing a data driven manuals and this illustrates a new form of intellectual production with distributed work and decentralizing power. It is in fact a democratic ideal because there is no academic reward or financial credit in this involvement.

Global Philology Planning Seminar Report

We have just created a Bachelor of science in DH in Leipzig University that aims the integration of computer science with humanities work and we are also organising an open conference. It tends to gather disciplinary needs and specificity to determine the structures and organisations we need.

Europeana, in a sense, seems to be stuck at the metadata level for further integration. Full integration will only happen when data circulates, otherwise it is just metadata. My goal here is to understand better how we can collaborate better than we do and to share sustainable development to connect resources and components. We are trying to have a list of core services and every historical document should have these services applied to it, to allow good things to happen.
For example, we could see all the places that were quoted in a document. All the names, every word should be analyzed. You should have syntactic analysis for all the words, you should be able to align all different versions that you just digitized and to compare them. It would be nice too to have text alignment across languages, ideally you could discover if there were translations of this work and align them. We could also produce on the fly lexicon for any word, use automated text mining, sentiment detection and lexicography to see patterns emerge.

Resources

- Global Philology Open Conference, Feb 2017: http://www.dh.uni-leipzig.de/wo/events/global-philology-open-conference/
- Open Greek and Latin as open data:
    - http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/
    - https://github.com/OpenGreekAndLatin
    - http://www.culingtec.uni-leipzig.de/ESU_C_T/node/379
- Patrologia Graeca:
    - https://mimno.infosci.cornell.edu/patgrec/
    - https://mimno.infosci.cornell.edu/
- Homer Multitext Project: http://www.homermultitext.org/
- Perseus Project: http://www.perseus.tufts.edu/hopper/
- CITE/CTS architecture:
    - http://www.homermultitext.org/hmt-doc/cite/index.html
    - http://cts.dh.uni-leipzig.de/
- Alpheios Reading tools: http://alpheios.net/
- Epidoc: https://sourceforge.net/p/epidoc/wiki/Home/
- Book alignement
    - http://books.cs.umass.edu/mellon/alignment.html
    - http://books.cs.umass.edu/beta-sprint/Demonstration/Demonstration.html
- Transkribus Project: https://transkribus.eu/Transkribus/

Contact

Gregory Crane is an Alexander von Humboldt Professor of Digital Humanities at Leipzig University. He is a specialist in classical philology and computer science.
He completed a doctorate in classical philology at Harvard University and worked as an assistant professor. He has the reputation of being a pioneer of digital humanities due to his development of the Perseus Digital Library, a freely accessible online library for antique source material. He was associate professor at TUFTS University and is now Winnick Family Chair of Technology and Entrepreneurship. He has received, among other awards, the Google Digital Humanities Award 2010.

Institutional website: http://www.dh.uni-leipzig.de/wo/gregory-crane/
Email: crane@informatik.uni-leipzig.de