NeDiMAH Working Group #4 Workshop

Open Digital Humanities: use and reuse of digital data in the Humanities

Amsterdam, 17 January 2013

Summary

The working group #4 of NeDiMAH focuses on Developing Digital Data. Building digital collections of data for research involves consideration of the subsequent use, and re-use, of these collections for research using ICT tools and methods. The use of digital collections for research has an impact on the creation, management and long-term sustainability of digital data, and the use of digital resources for the creation and publication of new knowledge is a vital part of the digital life cycle. The creation, description and structuring of digital resources, access to these resources (both in term of authorization and interoperability), their curation and preservation, and the publication of results of research they enable are connected parts of this process.

On January 17 2013, the working group hold a workshop entitled « Open Digital Humanities: use and reuse of digital data in the Humanities » in conjunction with the 8th International Digital Curation Conference (http://www.dcc.ac.uk/events/idcc13) in Amsterdam. The workshop addressed access and re-use of digital data in the arts and humanities and, more specifically, how to ensure findability of data, how to ensure trust and quality of data, how to ensure interoperability of data and how to ensure preservation of data.

The working group invited 5 speakers (William Kilbride, Laurents Sesink, Joris Pekel, Nuno Freire and Johan Oomen) and selected 2 additional speakers through a call for papers (Dominic Forest and Andreea Popa).

Scientific content

William Kilbride

William Kilbride's presentation reflected on the relationship between the arts and humanities and digital preservation. It noted that scholarship in the humanities is always to some extent historical and that research outputs therefore remain current for longer periods than might be the case with other disciplines. This creates a distinctive need for preservation. A series of standards have emerged in the last decade which codify expectations of best practice on how to preserve born-digital and digitized content and a growing digital preservation community has coalesced around issues of longevity, authenticity, and digital stewardship. This community has made rapid progress by borrowing tools and ideas promiscuously from sectors as diverse as forensics, engineering, conservation and industry. But the pace of development is itself a barrier to adoption: the glut of tools, standards, projects and services which the digital preservation community has generated can seem bewildering. More worryingly there is an abiding risk that the changing practices of the digital humanities mean that preservation tools are a poor fit to the actual needs of sector. His paper introduced some of the major and familiar themes from digital preservation in the last decade to

examine whether they are properly tailored to the needs of the humanities. It observed that digital humanities and digital preservation need each other: and that there is an on-going need for frequent and clear communications between sectors to make sure that digital research infrastructures are able to provide the sorts of preservation services that digital humanities require now and in the future.

Joris Pekel

Joris Pekel is from the Open Knowledge Foundation, a non-profit organization committed to the promotion of open content, with a strong focus on open data. He focused his talk on opening data in the Humanities, claiming that digital technologies have the potential to make knowledge available to anyone. He argued that open knowledge sharing through network technologies could bring up to date the principles of the 17th – 18th centuries Republic of Letters, by allowing frictionless ideas circulation in a global intellectual community.

Joris Pekel presented several tools and projects: the OpenGLAM project, "an initiative run by the Open Knowledge Foundation that promotes free and open access to digital cultural heritage held by Galleries, Libraries, Archives and Museums"[1]; Timeliner, a tool to buid timeline from google spreadsheets; crowdcrafting, an open source platform for crowdsourcing projects; textus, a platform enabling user to "collaboratively annotate texts and view the annotations of others, reliably cite electronic versions of texts, create bibliographies with stable URLs to online versions of those texts"[2]; DM2E (Digitized Manuscipts to Europeana), a project to provide Europeana with digital content (in particular digitized manuscripts) and to build tools to connect this content and enable its re-use in the context of Europeana Linked Data.

Nuno Freire

The talk given by Nuno Freire was actually co-authored with Valentine Charles and Antoine Isaac, all from Europeana and the European Library. This talk explored the Europeana enable the use of cultural heritage objects for digital humanities. Nuno Freire first presented how Europeana aggregate data provided by cultural institution all over Europe and how Europeana republish metadata in various way (such as through its portal or through linked data). He then presented the European Library which aggragate content from european national libraries and feed Europeana with this content.

Both Europeana and the European Library are committed to provide access to the data they host. For that purpose, they propose search API and invest Linked Open Data. The expected benefits are driving web traffic to data provider's websites, making Europeana and its partners become authorities for cultural heritage data and positioning Europeana and The European Library as a data hubs.

Nuno Freire presented three projects towards enabling the use of cultural heritage objects for digital humanities : Europena Cloud, which builds up on the Europeana infrastructure to make cultural heritage materials available for research by offering cloud infrastructure for data and contents, a licensing framework for reuse of content and a new research platform: Europeana Research; CENDARI (Collaborative European Digital Archive Infrastructure) which provides access to existing archives and resources in Europe for the study of medieval and modern European history; and DM2E, also mentioned by Joris Pekel.

Sesink Laurents

Laurents Sesink presented the Data Archiving and Networked Services (DANS). DANS is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands

1 http://openglam.org/
2 http://textusproject.org/

Organization for Scientific Research (NWO) created in 2006. Its mission is to promote and provide permanent access to digital research. For that purpose DANS has developed and maintain the narcis.nl portal which gives access to publications, datasets and description of researchers, research projects and research institutions.

The DANS digital repository implements a pragmatical policy, data being “open if possible, protected if necessary”. DANS pays particular attention to maintain its digital repository trustable and is actively involved in the certification standards establishment.

Johan Oomen

Johan Oomen heads Research and Development at the Netherlands Institute for Sound and Vision (NSIV), the Dutch audiovisual archive. In his talk, he argued for the need of more open, smart and connected audiovisual archives and identified three possible directions. First, Johan Oomen pointed the need for tools to dig into data (he reported the CISCO evaluated that it would take 6 million years to watch videos that will cross IP networks in a month by 2016). the Netherlands Institute for Sound and Vision has participated to the development of tools for facetted search (CoMeRDA), speech to text technologies, named entities and mashup (OpenBeelden), link entities with concepts of dbpedia (PoliMedia). The second direction is interoperability. Johan Oomen described the UEScreen project which gathers 27 european partners (including 19 audiovisual archives) and has provided 35.000 TV related items to Europeana. The Challenges faced by the project are content selection (quality rather than quantity), interoperability (archives used different metadata schema and video encoding standards), copyright, multilingality, interaction design. Finally, the last direction is crowdsourcing, i.e. using user participation for video content annotation.

Forest Dominic

Dominic Forest noted we have observed over the last ten years a considerable increase in the number of initiatives to digitize and make available on the web the information assets of the various branches of knowledge. In the scientific field, such initiatives to promote and disseminate knowledge lead to the development of "cyberinfrastructures", that is to say, new environments for the dissemination of research in the various fields of knowledge. In some targeted areas (including biology and biomedical engineering), the consequences of digitization initiatives have a direct impact on the development of software applications to assist research, analysis, structuring and management of information. Thus, these initiatives to digitize information resulted in several research projects whose aim is to assist the discovery and structuring of new knowledge from documents. Yet, despite all the interest in digitizing the information assets in the humanities (which is primarily textual in nature), very few projects have sought to develop and validate the performance of new approaches to exploit the full potential documentation offered in digital format. In a context where it is now impossible for a researcher to consult all the literature in a field of knowledge, it is therefore essential to develop tools that aim to assist the analysis and interpretation of textual documents.

In this presentation, he demonstrated how unsupervised text mining techniques can be used to assist information extraction, organization visualisation of a corpus of scientific documents in the humanities. The data that we have processed in our research is from the Erudit research dissemination platform (www.erudit.org). The approach he conducted is based on a generic methodology in data mining (Fayyad et al, 1996). This approach consists of four main steps:

- Step 1. The pre-processing of documents and the extraction of discriminant features.
- Step 2. The digital transformation of textual data
- Step 3. The application of text mining algorithms and the extraction of characteristic knowledge terms

- Step 4. The evaluation, visualization and interpretation of extracted information.

As part of his approach, Dominic Forest mainly puts to use the structural properties of clustering algorithms. He links the results of the clustering process to and information visualization process. The operating conditions of information visualisation is key, because it can deliver a user-friendly way to analyse the results of text mining process, which are traditionally presented in forms that are difficult to interpret (lists, contingency tables, etc.).

Popa Andreea

The new definition of Landscape (related to the European Convention on Landscape) "as a strategic resource for sustainable local development", helps to take account all types of landscapes that redefines the identities and roots of local Communities, giving them the chance to build sustainable local development based on natural, human and cultural resources. This new definition reiterate landscape in all related study sectors, which in present approach landscape problematic based on methods and techniques specific for each domain (ecological, geographical, planning, economical development, socio- cultural, historical, sociological, etc.).

In accordance to each European country practice, landscape is preserved and valuated on planning policies based on the sets of data generated by the primary study domain. Even if at European level is searched a common development policy concerning territorial development and landscape as subsidiary domain, the multitude of digital data and techniques used in order to asses landscape dynamic are not inter- correlated. At the end of the process, use of a specific set of data, imprint to local planning and development practice guidelines to protect and develop some structural elements of landscape, the whole as territorial system being continuously affected (consumed or totally transformed) by contextual changes (socio- economical, cultural and political inputs).

One of the problems arising is that various digital data sets are used in different ways with different result on long term policies, pointing the need to integrate the data sets used in this domain (recurrent from urban and territorial planning practice and geographical- ecological approach).

Andreea Popa specifically addressed the following points :

- The digital data sources for landscape dynamic assessment (practices at European level in accordance with primary study domain): GIS assessment, indicators data sets, photography, aerial photos and ortophotos comparison, audio and visual recordings, large data text collections for historical landscapes and reconstruction of lost landscapes- findability and use of sources, trust and quality of data
- The type of indicators used in order to asses landscape characteristics, landscape dynamic and development problematic
- The possibilities to integrate different niche targeted (sector) methods used to asses landscape in order to achieve a comprehensive approach and to ensure interoperability of data
- The possibilities to use and quantify non- technical data (humanistic and art resources) in a landscape dynamic matrix (lost landscape reconstruction, virtual landscapes, specificity and atmosphere quantifying- sensorial and sensitive aspects of landscape)
- The possibilities to develop a trans - disciplinary indicators system in order to asses and manage landscape dynamic (dynamic indicators as key issues in planning practice- real-time adaptive system of measures).

Final Program

- Kilbride William, Digital Preservation Coalition
  - "What we've learned about Digital Preservation and Digital Humanities: emerging practice (good and bad)"
- Joris Pekel, Open Knowledge Foundation
  - "The Digital Commons and the Republic of Letters"
- Nuno Freire, Valentine Charles & Antoine Isaac, Europeana / The European Library
  - "Europeana and Research: Enabling the Use of Cultural Heritage Objects for Digital Humanities"
- Sesink Laurents, Data Archiving and Networked Services
  - "Use and Trust"
- Oomen Johan, Nederlands Instituut voor Beeld en Geluid
  - "Towards more open, smart and connected audiovisual archives"
- Forest Dominic, University of Montreal
  - "Text mining, topic modeling and information discovery in cyberinfrastructures"
- Popa Andreea, University of Arhitecture and Urbanism Ion Mincu Bucharest
  - "Use of digital data in landscape planning: trans-disciplinary approach"

Participants

ANTONIJEVIC Smiljana, Penn State University

CAHOY Ellysa, Penn State University Libraries

DAY Michael, Digital Curation Centre

ENGEL Thomas, University of Applied Sciences Mainz

FANIEL Ixchel, OCLC Research

FOREST Dominic, University of Montreal

FOULONNEAU Muriel, Centre henri tudor

FREIRE Nuno, Europeana / The European Library

FREY Jeannette, BCU Lausanne

GRANT Rebecca, Digital Repository Ireland

GUERCIO Mariella, University of Rome La Sapienza

GUIBAULT Lucie, University of Amsterdam

GUY Marieke, DCC

HALBERT Martin, University of North Texas

INNOCENTI Perla, University of Glasgow

JOHNSTON Lisa, University of Minnesota

JORDAN Mark, Simon Fraser University

KILBRIDE William, Digital Preservation Coalition

KÖNIG Alexander, Max Planck Institute for Psycholinguistics

KOTARSKI Rachael, The British Library

LOTTER Lucia, Human Sciences Research Council

MAGUÉ Jean-Philippe University, of Lyon

MALLETT Elizabeth, The Open University

MINEL Jean-Luc University, Paris Ouest Nanterre La Défense

MORLOCK Emmanuelle, CNRS-TGE ADONIS

NOVOTNY Eric, University Libraries, Pennsylvania State University

OOMEN Johan, Nederlands Instituut voor Beeld en Geluid

PEKEL Joris, Open Knowledge Foundation

PHARO Nils, oslo and akershus university college

POPA Andreea, University of Architecture and Urban Planning Ion Mincu

RUDERSDORF Amy, Digital Public Library of America

RUMSEY Sally, University of Oxford

SESINK Laurents, Data Archiving and Networked Services

SKINNER Katherine, Educopia Institute

SMIT Frans, Gemeente Almere

THOMPSON Cheryl, University of Illinois

VAN WYK, Johann University of Pretoria

WALK Paul, UKOLN, University of Bath