Building and Developing Collections of Digital Data for Research

NeDiMAH Working Group #4 Meeting

Lyon, 1st March 2012

Scientific Report

The NeDiMAH project (Network for Digital Methods in the Arts and Humanities) is a Research Networking Programme that examine the practice of, and evidence for, advanced ICT methods in the arts and humanities across Europe, and articulate these findings in a series of outputs and publications. To accomplish this, NeDiMAH provides a locus of networking and interdisciplinary exchange of expertise among the trans-European community of digital arts and humanities researchers, as well as those engaged with creating and curating scholarly and cultural heritage digital collections. NeDiMAH maximizes the value of national and international e-research infrastructure initiatives by developing a methodological layer that allows arts and humanities researchers to develop, refine and share research methods that allow them to create and make best use of digital methods and collections. Better contextualization of ICT Methods also builds human capacity, and be of particular benefit for early stage researchers.

Activity of the NeDiMAH network is organized into seven working groups. The "Building digital collections" working group is justified by the fact that using of ICT tools and methods for research in the Arts and Humanities involves building collections of digital data. Use of such collections, as well as the anticipation of their subsequent reuse, raise many issues that affect each stage of the life cycle of digital data and that are addressed by the working group. In particular, the current and future diversity of tools requires consideration of interoperability constraints when describing and structuring the data. The

management of these data, their access, their curation and their long-term preservation require digital infrastructures enabling these operations. Access to these digital data also raises new legal issues. Finally, the role of these collections of digital data in the publication of new knowledge generated by research is yet to be specified.

The workshop that holds on March 1st 2012, was the first meeting of the "Building digital collections" working group and aimed at identifying the key topics to be addressed by the working group during the program. The organizer was Jean-Philippe Magué and the participants were:

- Mr. Bruno Bachimont, Compiègne, FR
- Mr. Tobias Blanke, London , UK
- Mr. Lou Burnard, Paris, FR
- Mr. Malte Dreyer, München, DE
- Mrs. Muriel Foulonneau, Luxembourg-Kirchberg, LU
- Mrs. Maria GUERCIO, URBINO, IT
- Mrs. Lucie Guibault, Amsterdam, NL
- Mrs. Perla Innocenti, Glasgo, UK
- Mr. Krister Linden, Helsinki, FI
- Mr. Nils Pharo, Oslo, NO

Scientific content

The 10 speakers, also members of the working group, represented as many different points of view on the use of digital collections for research in humanities.

Mariella Guercio

Mariella Guercio, with a background in archival science, defended the idea that common topics for qualifying digital data creation, use and keeping might be :

- Mapping of commonalities among disciplinary environments and sustaining an effort for a common or comparable (context-driven) terminology and for identifying parameters and metrics for evaluation
- Identification of usable existing services and standards
- Support for a standardized approach for processing capture, keeping and preservation of digital resources (life cycle or continuum models have to be planned and managed early and have to be compliant with existing standards)
- Definition of authenticity evidence as crucial component of any digital repository (both at creation and at preservation phase) by determining

standardized functions and developing automation processes for capturing and making available structured information and governing legal issues
- Development of advanced methods and tools for educating (at least in Europe thanks to the Bologna principles) qualified researchers both in disciplinary domains and in interdisciplinary environment

Krister Linden

Krister Linden, with a background in computational linguistics, argued for the sharing of best practices (for example by examining how things are done in the different countries), in particular when collecting digital data (with a focus on pricing and legal aspects) and when annotating digital data.

Nils Pharo

Nils Pharo, with a background in library and information science, pointed out the need for technologies allowing cooperation and interoperability between large organizations such as libraries or museums.

Malte Dreyer

Malte Dreyer, with a background in digital libraries, pointed out along discussing the topics to be addressed, the group should explicitly affirm a point of view. Dealing with digital collections involves not only the researchers working on the data, but also the organizations in charge of the infrastructures handling data and treatments. He suggested that issues might be perceived differently by the researchers and the infrastructures. He elaborated on the question of interoperability, arguing it has be thought not a the level of the data, but at the level the infrastructures. He also noted the need for distinguishing private and public workspaces, both needed at different stages of research: the former for the work of the researcher, the second for the data and the public dissemination.

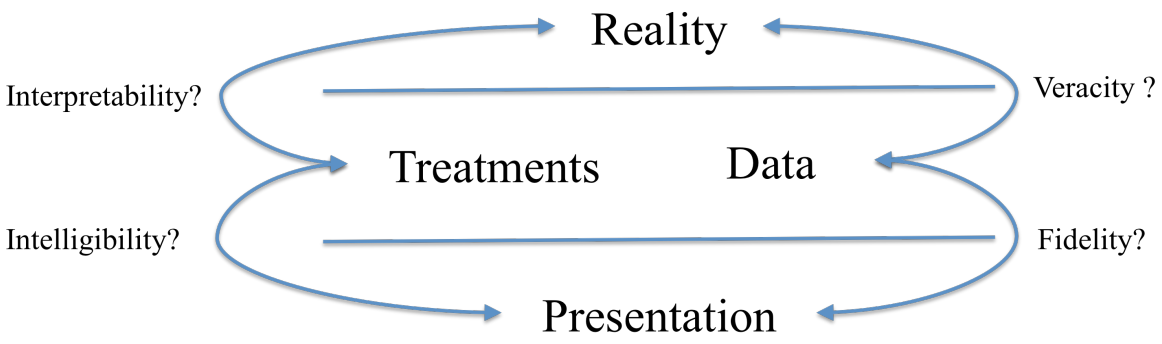
Perla Inocenti

Perla Inocenti, with a background in art history and digital cultural heritage, also drew our attention to the need for a common terminology. She proposed to focus on digital curation and preservation issues, such as risk assessment for digital libraries and digital repositories, interoperability between for digital libraries and digital repositories (at the organisational, semantic and technical levels), and Cross-domain collaboration models (networks, frameworks, policies)

Bruno Bachimont

Bruno Bachimont, with a background in philosophy and knowledge engineering, noted that we are experiencing a shift from documents to data. While documents are cultural construction, they are flattened into data where cultural, contextual marks are cancelled in order to build homogeneous collections that can be automatically exploited. He showed that this move introduces new questions: how to interpret the results of statistical and digital treatments applied on

flattened data? Do the results reflect model properties or data properties? How to show big databases or, more specifically, how to handle the tradeoff between showing something false but perceptible and interpretable, and something true but ununderstable? Given that data may be an invention depending on the way we collect them and that they may represent something that has never existed, what is behind the data and what are the data of? He proposed the following figure to summarize his thoughts:



Muriel Foulonneau

Muriel Foulonneau, with a background on research infrastructures for humanities, proposed three directions to be explored :

- Datasets: Which datasets are used for research purpose? How should they be formatted? How can we make that happen?
- Methodologies: Which methodologies for data collection in digital humanities?
- Tools: Are there shared tools that can be set up? Then which are the methodological aspects that need to be implemented?

Lucie Guibault

Lucie Guibault, with a background in law, oriented the group reflections toward legal issues, in particular regarding intellectual property protection (Publications vs. data, open access (green or gold), conditions of use of publication/data) and privacy issues (when individuals can be identified in research results)

Tobias Blanke

Tobias Blanke, with a background in computer science and philosophy, insisted on the need for methods to assure the transition from prototypes to sustainable tools.

Lou Burnard

Lou Burnard, with a background in humanities computing and based on his experience with XML TEI, argued for the necessity of standards :

- Because scholarship is increasingly about sharing resources (between humans as well as between machines),
- Because resources are increasingly seen as components to be integrated
- Because of the need to address preservation of digital resources

He drew the attention of the participants to the fact that standards may fail to to be accepted by a community when they are based on or express a theory that is not yet mature or when they are based on or express only one or a few of many contending or unresolved views in a particular domain. He thus argued that standards have to be community-owned, i.e. developed by the people who actually use them.