2009 Fifth IEEE International Conference on e-Science

# Alfalab

## Construction and Deconstruction of a Digital Humanities Experiment

Joris van Zundert
Huygens Institute (KNAW)
The Hague, The Netherlands
joris.van.zundert@huygensinstituut.knaw.nl

Douwe Zeldenrust
Meertens Institute (KNAW)
Amsterdam, The Netherlands
douwe.zeldenrust@meertens.knaw.nl

Anne Beaulieu
Virtual Knowledge Studio (KNAW)
Amsterdam, The Netherlands
anne.beaulieu@vks.knaw.nl

***Abstract*—This paper presents project 'Alfalab'. Alfalab is a collaborative frame work project of the Royal Netherlands Academy of Arts and Sciences (KNAW). It explores the success and fail factors for virtual research collaboration and supporting digital infrastructure in the Humanities. It does so by delivering a virtual research environment engineered through a virtual R&D collaborative and by drawing in use cases and feedback from Humanities researchers from two research fields: textual historical text research and historical GIS-application. The motivation for the project is found in a number of commonly stated factors that seem to be inhibiting general application of virtualized research infrastructure in the Humanities. The paper outlines the project's motivation, key characteristics and implementation. One of the pilot applications is described in greater detail.**

***Virtual research infrastructure; virtual research environments; VREs; virtual research communities; Alfalab, KNAW; collaboration; humanities***

## I. INTRODUCTION

It is within the e-science community a common place to state that the application of digital Humanities methods and techniques will create – and usually one might insert some superlative here – possibilities to explore new research questions. And certainly, there is evidence that this does happen, even in research domains where the use of computing, digital data and digital methodology is not common ground. [1] Another general statement often heard is that digital methods and computation are far from realizing their true potential for Humanities research. Although digital communication (e-mail, the web, chat, VOIP etc.) in general significantly transformed collaboration in Humanities research, certain forms of collaboration in which research groups are predominantly supported by dedicated and specialized digital research infrastructure are scarce. Arguably, it is this kind of collaboration that could yield significant new research results and methodological innovation [2] as well as contribute to agenda-setting in the Humanities. [3]

In the Netherlands, a new initiative called 'Alfalab' will create the setting and support for such collaborations in the humanities. Project 'Alfalab' will be active in three ways. First it will guide the development of a web-based virtual research infrastructure for collaborations in the humanities. Secondly, it will seek to avoid the shortcomings of other initiatives, as identified in a number of studies and reports [4] in order to support scholars in their explorations of such new tools. And third, project 'Alfalab' will critically consider the dynamics of expectations about new technologies for scholarship and interactions with potential users [4] [5].

## II. CRITICAL REFLECTION

Great expectations are rife around any new technology, expectations that go more often than not unfulfilled. At times, disappointment seems to come from the technologies that do not deliver what was promised. In other cases users are blamed for failing to see the potential of new tools. For example, the recent report by the American Council of Learned Societies (ACLS) Commission on Cyberinfrastructure for the Humanities and Social Sciences [6] signals a specific set of issues and identifies these as causes for a purported lack of uptake by users. In view of this, there might be several complex reasons why the Humanities in general do not seem to profit yet from virtualization of research networks. The commission suggested that, although there's no doubt about the usefulness of collaboration in the sciences, there are relatively few formal digital communities and relatively few institutional platforms for web-based collaboration in the humanities. Solitary scholarship and non-virtualized collaboration are still dominant. The report of the commission further indicated that probable causes for this situation were the specific nature of data in the Humanities (which is characteristically complex, heterogeneous and ambiguous) and the lack of critical mass of large quantities of such heterogeneous digitized data for digital and computational methodology to take off. We consider that a lack of proper and easy access to digital data, tools and expertise adds to the problem.

978-0-7695-3877-8/09 $26.00 © 2009 IEEE
DOI 10.1109/e-Science.2009.8

Embracing the arguments made by the ACLS-report and several studies that showed the situation to be rather similar in The Netherlands [7], The Royal Netherlands Academy of Arts and Sciences (KNAW) decided to shape a new initiative, in coordination with its own strategy for developing Humanities research [8]. In August 2008 a project proposal 'Alfalab' [9] was accepted by the Board of the KNAW and the project formally launched on March 1, 2009. It's first phase will last until March 1, 2011 and incorporates the contributions of at least five Humanities research institutes of the KNAW. Project 'Alfalab' aims to involve about the same number of research groups and institutions outside the Academy. The initial economic dimension of the first phase of the project is about 1.4 million Euros.

## III. ALFALAB AS A SANDBOX

In a nutshell Alfalab can be characterized as a sandbox to critically reflect on the ACLS's findings and assumptions. More elaborately Alfalab could be described as an exploratory experiment in collaboration on Humanities digital instrumentation engineering and dissemination. Collaboration is to be pursued on all dimensions: between the organizations and people building Alfalab; on the dimension of expertise and disciplines (in the first stage, across ethnography, science studies, historical textual research and geographical information sciences); on dissemination where the aim is to have both users and creators take advantage of each other's experience and expertise; and across actors involved with new infrastructures, whether as user, developer or decision-maker.

In its first stage Alfalab will create a small web-based setting that tries to satisfy the conditions the ACLS's and other reports specify for engaging Humanities researchers in the use of digital tools, data and supporting infrastructure. The foremost of prerequisites and Alfalab's approach to them will be outlined below.

Alfalab as a virtual research environment comprises a virtual laboratory, combining a tools registry and a dataset registry. In a visually attractive and straightforward web GUI a tool dataset broker agent will preselect suitable data sets once the user/researcher selects a tool, and vice versa. The well known TAPOR [10] portal might be taken as a 'role model' for this part of the research environment. The convenience of having data sets presented along with applicable tools should preempt the problem of readily availability of tools in this testing context.

Alfalab is an exploratory project, and will critically consider which tools to support. It will not try to incorporate all possible digital data and tools available within the Humanities in The Netherlands or abroad. Moreover: it doesn't have the financial dimension to be able to realize that. Rather two selected domains (historical textual scholarship and historical geo-spatial research) will form the focus for the registries and brokerage of digital data and tools. This small scale approach of the first phase is primarily meant to test the rather implicit assumption that such accessible and easy to use brokerage will lead to higher use and impact of digital tools and data. We will also test the extent to which our critical review of tools for inclusion, on the basis of their design and implementation philosophy, enables us to select viable options.

## IV. TOOLS FOR HETEROGENEOUS DATA

Two teams of combined Humanities researchers and IT engineers will develop a small number of demonstrator tools. These tools mainly focus on the problem of heterogeneous data on a proof of concept basis. Specifically they will show how, rather than being impeded by heterogeneous data, the tools' functions may be geared around common tasks. The implementation of the tools and the integration of these into the virtual research environment will also serve to demonstrate Alfalab's philosophy of distributed services. The principle of distributed services allows for multiple instantiations of the same tool in different virtual environments. This means that the same service might both be served through the Alfalab virtual research environment for convenient discovery, as well as within the website of the originating institution. This would also allow tools to be individually branded by the developing and maintaining institution, thereby addressing issues of credit and reputation.

Currently two demonstrators are being developed. One is dubbed 'TextLab' and comprises a number of digital tools for textual exploration. To name some: a named entity recognition (NER) service for onomastic studies based on a computational approach; transcription and annotation of textual sources; inductive support for auto-suggesting textual annotation. The latter tool is modeled after social network suggestions known from such sites as LinkedIn and Facebook. This tool will support researchers transcribing or annotating text with messages of the form "Other researchers annotating the words [....] added the following annotation [....]". Suggestions will be based on word form distance computation and context matching. Such a tool should thus show the feasibility of task driven services to be able to support digital scholarship even in contexts of heterogeneous and ambiguous data. The heterogeneity in this case is 'semi-solved' through word distance and context computation. The problem of ambiguous data is 'semi-solved' by the human-directed annotation suggesting algorithm. Such tools can greatly facilitate the repetitive tasks involved in the curation and enrichment of textual scholarly data, notwithstanding the problems of heterogeneity and ambiguity.

It should be noted though, that Alfalab does not aim to 'solve' the problem of heterogeneous data. It's even arguable whether there is such a problem that needs solving. Rather Alfalab is trying to tackle the problem of sustainability or applicability of tools tailored to heterogeneous data. Tools should therefore be defined as task-oriented service components. A task-oriented algorithm can be modeled completely separate from the form or format of the data it's aimed to operate on. Analyses of specific forms of data can be supported through providing specific importers and/or converters that curate or rewrite the data for the analytical model of the task algorithm. In such a way the inherent analytical value of an algorithm can be applied with relative ease to multiple heterogeneous data sets. This approach also has the distinctive advantage that it takes the particular conditions of humanities as a starting point, rather than trying to impose models from tools developed for other scientific domains [11].

The second demonstrator is called ‘GISlab’. It will display, as a part of its research and development track,

some of the principles of coping with heterogeneous and ambiguous data that were described in the former paragraph. The GISlab offers tools for (and tutorials on) applying Geographical Information Systems (GIS) to historical material for research purposes. An example of this is the accessible and easy to use web-based map annotation tool. Two research projects use this tool. Within the digital rural microtoponyms project of the Meertens Institute the tool is used to align ('georeference') maps that were annotated by hand, in the course of transforming the written annotations into digital representations for analysis. This project is presented as a case study in greater detail in the latter part of this paper.

The georeferencing of the microtoponym maps is problematic but, within certain limitations, possible (see the case study). In contrast the MAPS project run by Leen Breure (affiliated with the University of Utrecht and the National Archives, which are external parters of the KNAW and project 'Alfalab') uses digitized maps in a wide variety of cartography method and style [12]. To give but one examples: the corpus in use incorporates 17th century townscapes. The extent to which features of such historical maps in the broadest sense can be geo-referenced is rather smaller compared to the ones in the mycrotoponyms project. But annotation and categorization of annotations linked to specific coordinates within the image data can still be accomplished. This does not mean that annotations become less valuable, just that there's a larger uncertainty in their geo-spatial positioning. Yet another application of the image annotation tool is sought in connection to 'TextLab', where this tool can be used to annotate the visual 'landscape' represented by manuscript facsimile (referencing and identifying illuminations in the manuscripts for example). These examples demonstrate how task orientation can abstract away from the specificities of form and format, to better be able to grasp the interdisciplinary application potential of digital tools.

## V. SHARING KNOWLEDGE

Dissemination should address the problem of a lack of knowledge about existing digital tools, data, infrastructure and possible application – which is one of the other key aspects the ACLS report defined. Alfalab is in this respect geared both towards outward dissemination and inward feedback. Outward dissemination comes in the form of workshops, tutorials, digital manuals and documentation, as well as getting graduate students involved by way of introducing a course on digital methodology in university curriculum. Arguably more important is the feedback cycle Alfalab wants to create. Alfalab does not want to be yet another implementation of a top down vision on virtual research environments for the Humanities. Rather it's interested in producing a concise and comprehensive prototype of such an environment and the supportive infrastructure, to be reflected on by part of the Humanities community. 'Part' because the demonstrators and existing software that will be applied in the workshops are geared towards two specific subfields in the Humanities. Through these workshops Alfalab attempts to create, generate and evolve the momentum, modus operandi and the digital concepts that actually do support virtual collaboration communities.

In all aspects Alfalab is thus an exploration and experiment. It is applied research into virtual collaboration to further the use and engagement with digital data and instrumentation in the Humanities. And this starts at the very basis of the project. Alfalab is engineered and evaluated by a virtual Humanities R&D team. Digital instrumentation supports the project members in this virtual collaboration where individual partners are distributed over different locations and organizations within and outside the KNAW. Blogs, wikis, web-based issue tracking software, project dashboards, chat, VOIP, code repositories and mail are in place to provide all possible means of working together in a virtual project room. The Alfalab R&D team intends to practice what it preaches.

Collaboration extends on a further level as Alfalab is happy to share a number of basic tasks with other supportive academic partners rather than reinvent the wheel. CLARIN-NL as a provider of knowledge on standard solutions for data repositories, meta-data formats, authorization et cetera is meanwhile connected to Alfalab. The value in this for CLARIN is that project 'Alfalab' will in a sense test whether the good and preferred practices CLARIN suggests are indeed also viable in Alfalab's context [13]. The knowledge and experience that this will generate may be taken as feedback for future dissemination on good practices and preferred standards by CLARIN.

Finally, and crucially, collaboration is extended towards the community of Humanities researchers. A call will be published for researchers to participate in a number of workshops, in which they can apply existing software to their digital research data as well as explore tools that have been developed within the scope of Alfalab. The interactions at these workshops will be documented and analyzed. The feedback generated will be used to further tailor the tools and dissemination framework of Alfalab. Dissemination of the tools and applications will also be implemented through the university curriculum, in order to strengthen the chances of an active research community evolving around Alfalab.

The overall process, its implications, impact and success factors will be studied and evaluated through ethnographic and survey research by the Virtual Knowledge Studio for the Humanities and Social Sciences, an institute of the KNAW [14]. This research will result in a comprehensive study on aspects involved in the creation of rich and sustainable research practices using digital tools and distributed infrastructure.

## VI. DIGITAL RURAL MICROTOPONYMS

As stated above, one of the pilot projects of the GISlab involves the collection of rural microtoponyms of the Meerens Institute [15]. The Meertens Institute, an institute of KNAW, studies the diversity in language and culture in the Netherlands and onomastic variation is part of this. The onomasticians at the Meertens Instituut specialize in the study of proper names. This onomastics discipline consists basically of two sub domains. The first sub domain is called anthroponomy. It is the study of personal names. The second domain is called toponymy. That is the study of place names. Apart from research, the Meertens Institute also concerns itself with documentation and

providing information to third parties in the field of Dutch language and culture. The Institute possesses a library and a vast documentation system, of which databases are a substantive and crucial part [16].

The collection of rural microtoponyms used in the GISlab pilot has been building up by the Meertens Institute since 1948. For thirty years, data about rural microtoponyms in the Netherlands has been gathered. This collection is the largest onomastic collection at the Meertens Institute [17]. Rural microtoponyms is the collective term for the names of small entities in both natural and man-made landscape. The first category covers all sorts of rugged features, such as moors, natural forests and marshes, as well as streams, lakes etcetera. The second covers cultivated landscape and includes individual parcels as well as arable land, grazing land and man-made forests. Examples are for instance 'Hogeweid' (High field), 'Molenstuk (Mill piece) and 'de Punt' (the Point) [18]. The collection of the Meertens Institute comes mainly on handwritten cards which state the name, the origin of the name, the location and the soil composition and use. Some of the cards also contain written information on the back. The collection contains an estimated 240,000 microtoponyms on individual cards and over 2400 topographical maps of various origin upon which the microtoponyms are marked. Another important source of microtoponyms is the collection of the Fryske Akademy. This Akademy is also part of Alfalab and its collection of digital microtoponyms is substantial [19].

Microtoponyms are not only an excellent source of information for onomasticians, they are also a focus of interest for other disciplines. In 2003 a workshop was organized at the Meertens Instituut on the study of microtoponyms in the twenty-first century. Presentations were held by, amongst others, Prof. Drs. J. Vervloet from the Socio-spatial Analysis Group of the Wageningen University and Prof. Dr. H. Mol from the Department of History of the Fryske Akademy. The conclusion was that if the microtoponyms could be digitized with the aid of a geographic information system, it would facilitate and open up new avenues of research in various other academic fields such as geography, archeology, variation linguistics and history [20]. With this conclusion in mind a pilot project was started in 2004 called: DIgitization of rural MIcroTOponyms (DIMITO). The key objective was to explore the potential for digitization on the basis of a small sample from the available material and to find out whether is it useful and feasible to digitize the entire collection [21].

In 2006 the project was completed. A working prototype with the microtoponyms of the municipality of Heiloo (400 microtoponyms and 9 maps) was constructed [22]. The conclusion was that, even though some aspects would present a problem, it is possible to digitize the microtoponym collection and to set up a geographic information system. One serious obstacle is the determination of the location of microtoponyms, despite the presence of maps. Many of the microtoponym maps are decades old and have very little affinity with modern digital maps. Old maps often have imperfections which are copied to the digital environment, and hence, to the location of the microtoponyms [23].

In order to keep the presentation and accessibility of the original material as good as possible, it was decided to link each microtoponym to a scanned image of the card containing the data and of the map. This means that (good quality) digital copies of the original material can be consulted for each microtoponym. The database can also be accessed from other perspectives, besides the geographic. It contains the data from the cards and other fields that can be filled thanks to the available material. The digital source that has been developed does not only provide easier access to the collection, it opens up opportunities for new research questions. Paper data can only be accessed through the cards; this data can be consulted in multiple ways. Qualitative, quantitative and geographic data can be requested displayed and exchanged [24].

After 2006 the success of DIMITO still had to be capitalized and a suitable framework, that would be innovative and practical at the same time, was needed. In 2009, Alfalab provides the excellent opening to take the next step in digitizing and using the collection. The Meertens Institute will scan all cards and maps and the GISlab will deliver a platform upon which the maps can be georeferenced, microtoponyms can be localized and information can be added and exchanged. It will also provide a possibility to tag information. The collection of the Fryske Akademy will be part of the pilot as well. The workshops that will be organized in the course of the Alfalab project will offer opportunity to researchers from in and outside the KNAW to evaluate the usability and applicability of the data and tools provide through the digital rural microtoponyms project [25].

## VII. CONCLUSION

Project 'Alfalab' is an initiative of the KNAW drawing five of its institutes into an exploration of aspects of virtual research collaboration and supporting infrastructure in the Humanities. The first phase of the project focuses on success factors for virtualized research collaboration. To this end virtual collaboration is implemented on all possible levels. It is implemented within the R&D needed for engineering Alfalab as an virtual research environment. It will be implemented as collaboration with and between researchers in the broader field who are the users of the virtual environment 'Alfalab'. Also collaboration will be established with other infrastructure oriented networks like CLARIN-NL.

In the course of the project a number of assumptions on virtual collaboration in the Humanities will be put to the test. To achieve this, project 'Alfalab' will implement two demonstrators or pilots on specific sub-domains in the Humanities in its first phase. By way of workshops and dissemination, the usability and applicability of the tools and the supporting digital infrastructure developed will be tested by Humanities researchers. The development of these demonstrators will be a key site for developing new research practices and developing new approaches to data.

In building and disseminating the demonstrators as well as in its other activities, Alfalab tries to meet the conditions and challenges identified in reports such as that of the ACLS. Project 'Alfalab' will attempt to validate the assumptions made by various such reports and studies towards the perceived current and future engagement of

digital tools and data in the Humanities. Based on the findings, a motivation will be developed for the second phase of Alfalab. This second phase should be geared towards enhancing Alfalab's relationship to humanities scholars, upscaling the number of tools, data services and partners involved. All this, in a mode and form that leverages the potential of digital instrumentation for scholarship [26].

In essence project 'Alfalab' is constructing a modest scale digital infrastructure to support a small number of specific virtual collaborative Humanities research endeavors. By de-constructing the project and its implementation, the researchers involved will be able to analyze and determine the success factors and shortcomings that will help the progress of virtual research collaboration and supporting digital infrastructure in the Humanities.

## REFERENCES


[1] D. Staley, Computers, Visualization, and History: How New Technology Will Transform Our Understanding of the Past. New York: 2002.

[2] O. Boonstra, L. Breure, P. Doorn, Past, Present and Future of Historical Information Science. Amsterdam: 2004.

[3] P. Wouters, "The agenda-setting role of e-research," in W.H. Dutton (ed.), World Wide Science: the promises, threats and Realities of e-Research. Oxford: Oxford University Press, unpublished.

[4] N. Brown, M. Michael, "A sociology of expectations: Retrospecting prospects and prospecting retrospects," in Technology, Analysis & Strategic Management, 15, 2003, pp. 3-18.

[5] P. Wouters, K. Vann et al, "Messy Shapes of Knowledge - STS Explores Informatization, New Media and Academic Work," in E.J. Hacket et al (eds.), The Handbook of Science and Technology Studies. Cambridge, Mass.: MIT Press, 2008, pp. 319-352.

[6] M. Welshons et al (eds.), Our Cultural Commonwealth, The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. ACLS, 2006. See also: D. Atkins et al, Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. NSF, 2003.

[7] W. Bijker, B. Peperkamp (eds.), Geëngageerde geesteswetenschappen, Perspectieven op cultuurveranderingen in een digitaliserend tijdperk. The Hague: AWT, 2002. See also: Wetenschap gewaardeerd! NWO-strategie 2007–2010. The Hague: NWO, 2006.

[8] Duurzame wetenschap, Strategisch plan KNAW 2007-2010. Amsterdam: KNAW, 2006. See also: Th. Mulder, Strategische Visie op de Instituten van de KNAW, concept 2. Amsterdam: 2008, unpublished.

[9] Alfalab, Aanvraag voor de opbouw van een digitale onderzoeksinfrastructuur voor de Geesteswetenschappen – de eerste fase, Koninklijke Nederlandse Akademie van Wetenschappen. Asmterdam: 2008, unpublished.

[10] Cf. http://portal.tapor.ca/portal/portal accessed, June 30th, 2009.

[11] P. Wouters, A. Beaulieu, "Imagining e-science beyond computation," in C. Hine (ed.), New Infrastructures for Knowledge Production: Understanding E-Science. Idea Group, 2006, pp. 46-70.

[12] Charles van den Heuvel: "MAPS: Manuscript map Annotation and Presentation System - Linking formal ontologies with social tagging to (re-)construct relationships between manuscript maps and contextual documents," in Kate Singer (Ed.), Digital Humanities 2009, Conference Abstracts. MITH, 2009, pp. 138-141.

[13] Cf. http://www.hum.uu.nl/clarin-nl/relatedprojects.html (accessed, June 30th, 2009)

[14] Cf. http://www.virtualknowledgestudio.nl/projects/alfalab.php (accessed, July 8th, 2009)

[15] Alfalab, Aanvraag voor de opbouw van een digitale onderzoeksinfrastructuur voor de Geesteswetenschappen – de eerste fase, Koninklijke Nederlandse Akademie van Wetenschappen. Amsterdam: 2008, pp. 26, unpublished.

[16] Cf. http://www.meertens.knaw.nl, htttp://www.knaw.nl and http://www.naamkunde.nl (accessed, June 30th, 2009).

[17] Cf. the archive of the Meertens Instituut, collection no. 49, collection of field name maps ca. 1860 – 1964, the Archief Meertens Instituut, collection no. 99, fieldname collection 1941 – 1992 and the Field Name Map Database by Leendert Brouwer from the Meertens Instituut.

[18] Ibidem, see also: M. Schönveld, Veldnamen in Nederland (Amsterdam, 1950) and R. Rentenaar, "Plaatsnamen in historische bronnen," in Naamkunde 2. Leuven: 2002, pp. 137-148.

[19] Cf. http://www.fryske-akademy.nl and http://www.hisgis.nl (accesed, June 30th, 2009).

[20] H. Beijers (et al.), Veldnamen als historische bron, een handleiding voor methodisch onderzoek ('s-Hertogenbosch, 1991). H. Bennis, 'Naamkunde als discipline', in: Naamkunde 2 (Leuven, 2002, pp. 129-136). See also: http://www.meertens.nl/books/veldnamen. (accesed, June 30th, 2009).

[21] D. Gerritzen, Veldnamen in Noord-Nederland. Een pilot voor een multidisciplinaire database. Subsidy application for the Digitization Fund. (unpublished, Amsterdam, 2003).

[22] D. Zeldenrust: prototype DIMITO. accessed, June 30th, 2009.

[23] D.A. Zeldenrust, "DIMITO: Digitization of rural microtoponyms at the Meertens Instituut," in Humanities, Computers and Cultural Heritage. Amsterdam: 2005, pp. 301-307 and D.A. Zeldenrust, "Digitalisering van microtoponiemen van het Meertens Instituut," in Naamkunde 2. Leuven: 2008, pp. 121-133.

[24] Ibidem.

[25] Alfalab, Aanvraag voor de opbouw van een digitale onderzoeksinfrastructuur voor de Geesteswetenschappen – de eerste fase, Koninklijke Nederlandse Akademie van Wetenschappen. Amsterdam: 2009, unpublished.

[26] Ibidem, pp. 26.