If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities

Joris van Zundert [*]

Abstract: *»Wenn Ihr baut, werden wir wohnen? Digitale Infrastrukturen großen Maßstabs als Sackgasse für die Digitalen Geisteswissenschaften«*. Programs aiming to develop large scale digital infrastructure for the humanities motivate this development mostly by the wish to leverage methodological innovation through digital and computational approaches. It is questionable, however, if large scale infrastructures are the right incubator model for bringing about such innovation. The necessary generalizations and standardizations, management and development processes that large infrastructures need to apply to cater to wholesale humanities are at odds with well-known aspects of innovation. Moreover, such generalizations close off many possibilities for exploring new modeling and computing approaches. I argue that methodological innovation and advancing the modeling of humanities data and heuristics is better served by flexible small-scale research focused development practices. It will also be shown that modeling highly specific distributed web services is a more promising avenue for sustainability of highly heterogeneous humanities digital data than standards enforcement and current encoding practices.

Keywords: digital humanities, infrastructure, tool building, sustainability, innovation, heuristics, methodology.

Introduction

I think it is paramount that we as scholars in the humanities, regardless of whether we are applying digital and computational approaches or not, understand why big institutionally-based digital infrastructures are a dead end for information technology development and application in the humanities. This message is urgent and essential, for if we do not grasp it we stand the risk of wasting grand effort and funding in the near future on delivering empty infrastructures bereft of useful tools and data. In the middle of a small-scale-focused, multi-faceted, patched-together, interconnected, very slow but ever developing technological humanities landscape, these tall big bulky structures will be waiting for a horde of uniformly behaving humanities scholars that will

[*] Address all communications to: Joris van Zundert, Huygens Institute for the History of the Netherlands, Royal Netherlands Academy of Arts and Sciences, PO Box 90754, 2509 LT The Hague, Netherlands; e-mail: joris.van.zundert@huygens.knaw.nl.
I owe many thanks to Tara L. Andrews, Ronald Haentjens Dekker, and Gregor Middell for their invaluable comments without which this article could not have been written.

never come. These infrastructures will be like the disastrously wrongly planned and developed highways that connect nothing to nowhere.[1]

There are many directions from which we can approach the question of why big all-encompassing all-serving digital infrastructures are meaningless and useless for digital humanities technology development. In fact there are so many aspects to consider that making the argument is hard, making it akin to that feeling of the whole thing being so wrong on so many levels and convoluted in so many ways that it is hard to figure out where to start explaining. I will narrow my argument to what I can say on the subject based on the experience and knowledge on innovation of technology and methodology for digital scholarly editions that I accumulated during the past five years running the Interedition initiative (Interedition 2012). This knowledge touches, amongst other themes, upon processes of innovation, changing research practices, software development and sustainability. Although this knowledge and experience predominantly pertains to the creation, functioning, analysis and preservation of digital scholarly editions and related tools and data, I do think the argument, mutatis mutandis, applies to the relationship between big digital infrastructures and the majority of digital scholarship in the humanities.

The Innovation Aspect

Given recent large investments in projects such as BAMBOO (<http://www.projectbamboo.org/>), DARIAH (<http://www.dariah.eu/>), and CLARIN (<http://www.clarin.eu/external/>), there seems to be a certain consensus among funders and policymakers that there is a real need for the humanities to shift its methodology into the digital realm. The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences, for example, heralds digital and computational approaches as drivers of methodological innovation in humanities. Development programs on which large-scale infrastructure projects like BAMBOO, DARIAH, and CLARIN are based adopt similar policies, terming the humanities as inherently too conservative to adopt an innovative digital methodology. It is also routine to point out a lack of proper formalization for the field, due to the ephemeral and heterogeneous nature of humanities research. Furthermore, there is an absence of a critical mass of digitized research data. Insufficient IT training and skills are furthermore to blame. (Welshons 2006) In all, true digital methodological innovation will not happen without considerable additional impulses to develop tools and infrastructure.

[1] For a particular daunting example of real world infrastructure planning failure, see <http://www.baltimorebrew.com/2011/02/01/highway-to-nowhere-shut-down-and-baltimore-doesnt-notice/>.

Thus, large humanities digital infrastructure projects are focused on snowballing IT-based methodological innovation into a humanities domain-wide push. These large infrastructures promise to deliver a shared digital infrastructure to store and sustain research data in such ways that the data will be uniquely identified, discoverable and usable. Appropriate tools to curate, discover, analyze and visualize research data will be developed, maintained, and made usable for humanities scholars. These tools will be generalized and intuitive and will thus usher in the innovation of methodology by making pattern mining, mass data analysis, and visualization of the future mass of digitized humanities data tractable for the typical humanities researcher (Clariah 2012).

But if methodological innovation based on information technology is the key to the future of the field, do big structures, either organizational or digital, form the right incubator pattern? First of all one must critically assess whether a revolutionary paradigm shift must necessarily be tied to a shift towards the digital. The primary purpose of digital technology is not to shift paradigms. Rather, digital technology is most often simply used to enlarge the efficiency of existing processes, for instance through automation (Haentjens Dekker 2011) and distribution of workload (Beaulieu 2012). And although computational tools may well add to a methodology, they should also warrant that existing heuristics and hermeneutics are appropriately translated into their equivalent digital counterparts, especially in a field where heterogeneity of data and multifaceted approaches are not regarded as reducible noise but as essential properties of the research domain. The focus of the big infrastructures under development on a revolutionary paradigm shift and on large scale generalized tools and data seems at odds at least with the precautionary principle of 'first do no harm'. What is perceived as a conservatism in a field may actually be the justifiable argument that IT has not produced very usable or even useful tools for humanities scholars until now (Van Zundert 2012; Gray 2010, 47). Furthermore, given that the library world is now digitizing on a massive scale (KB 2010), it is maybe not the critical mass of digitized data that is missing – though it is debatable enough whether the quality of its web availability is appropriate – but rather the tools that would allow humanities scholars to use those digitized materials in any useful way for research.

It is unlikely that the large scale digital infrastructure projects will solve this problem of lack of tools any time soon, just by being infrastructures of impressive scale. Mass data analysis tools have not been developed through big digital infrastructures until now, but through industry and local university effort. Moreover as Gregory Crane puts it: "[d]ocument analysis, multilingual technology, and information extraction can be modeled in general terms, but these technologies acquire meaning when they are aligned with the needs of particular domains" (Crane 2006). Aligning with the needs of particular domains means adapting and applying general model digital tools, or even developing purpose built tools, to very specific domain constraints problems. It is at this

very specific level where methodological innovation happens based on IT capabilities – not as a general principle but as a specific solution to a specific demand.

Indeed we need to be very realistic about the level, dimension, and impact of IT on methodological innovation in the humanities. Although the recent debate (Ramsay 2011; Fish 2011) surrounding the MLA (<http://www.mla.org/>) may suggest an influx of the digital into humanities, the actual contribution of digital humanities to innovative tools and infrastructure is very modest. As Ramsay points out, writing a blog post is essentially not a methodological innovation, just as using email or a text editor cannot be really called a leap of innovation. Application of conventional interpretative frameworks to digitally born culture, such as a critical examination of online literary reception (Boot 2011), borders on innovation in the sense that it broadens the field of study. But such research is not innovative in the sense that heuristics change or are renewed. Tool building in itself certainly does not necessarily constitute innovation. A bibliography is a bibliography. The fact that it is digital does not represent innovation in itself, though it is of course highly useful and convenient to finally have such resources digitally available.

The expression of heuristics through tools and addition of new heuristics or support for interpretation through construction of models and analytical tools does constitute an act of innovation (Rockwell 2012). But such developments are exceptional. Of course there are examples past and present; we can point to network analysis of correspondences (<http://ckcc.huygens.knaw.nl/>) or automated collation (Haentjens Dekker 2011) as projects that truly do touch upon new heuristics. But all in all, real innovation, i.e. methodology changing innovation, based on information technology or computational approaches seems to be rather scarce in the humanities. The larger part of digital humanities seems more concerned with digitization, perhaps, than with methodological change. True innovation is a niche in a niche within digital humanities. To support and foster that with large scale infrastructure seems excessive.

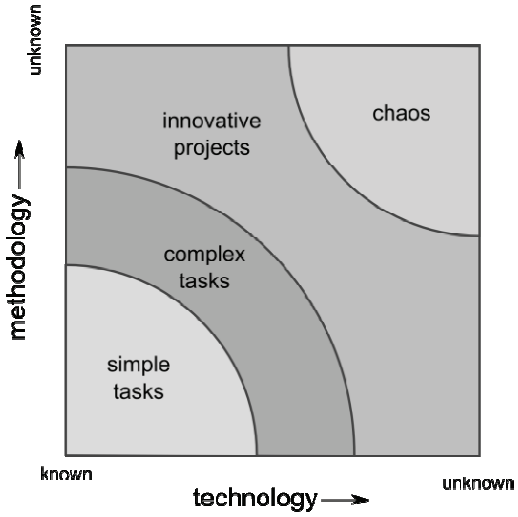
The Generalization Paradox

Digitization is a well-tried and tested terrain for activities meant to generate digital resources for humanities research. To a certain extent the types of workflows, tasks, and technical properties involved in the digitization of research material are well known. There is an argument, therefore, that large scale digital infrastructure facilities are useful for supporting library institutions in digitizing, and that their common infrastructure serves as a natural host for digitized data in order to expose it for use. However, the technical and support needs vary so much from library to library institution that they do not seem easily generalizable. Moreover, a considerable number of libraries are digitizing or even already have digitized their collections, apparently without the help

of large scale humanities digital research infrastructures. And although digitization is of course of pivotal importance to the humanities, it does not constitute humanities research. For instance, IMPACT (<http://www.impact-project.eu/>) is a project originating from the digital library world that spurred technical research into improved OCR. Thereby it yielded new tools through pushing the boundaries of OCR technology. However, in itself the project did not result in new humanities endeavors. The application of the IMPACT tools does not seem to be the type of computational endeavor that big infrastructure programs point to as paradigm shifting application of IT to the humanities domain.

If it is already hard to see how relatively straightforward digitization projects are to be supported in a generalized way, it is nearly impossible to establish what a generalized infrastructure would look like for high-end innovative projects geared towards humanities research – the sorts that involve experimental pattern detection, large scale analysis of noisy data, and exploratory knowledge visualizations. This near-impossibility follows from the experimental character of the research. The uncertain and volatile nature of innovation determines that it is hard to establish the forms and requirements of any underlying technology or infrastructure (cf. fig. 1).

Figure 1: Innovation Matrix, Adapted from (Stacey)



Innovation by definition is the exploration and investigation of that which is unknown by doing and experimentation. Thus, if the large infrastructural projects are concerned with innovation, the question that they must pose to themselves is: how do we deliver an infrastructure for something that is unknown? And how do we cater to unknown research questions?

Having already concluded that there are only very few true methodological game-changers in the application of IT to the humanities, we must now also conclude that we cannot be sure about the generalizability of the technological makeup of any existing game-changers. For potential game-changers it is far too early to tell what big infrastructure would support their volatile technology usefully and indefinitely. The Circulation of Knowledge project executed at the Huygens ING in the Netherlands can be held up as just one example of how volatile innovative projects can be. Its key objective is the development of tools to trace the evolution of ideas and concepts over time in learned correspondences from the time of the Dutch Republic. This requires, for example, pushing the state of the art of topic modeling over multiple languages barriers and phases. In the original project plan several technologies such as LSA (<http://en.wikipedia.org/wiki/Latent_semantic_analysis>) and Emdross (<http://emdros.org/>) were regarded as key technologies to implement the wish for temporal cross-language topic modeling. Meanwhile the combined expertise of humanities and information technology researchers has moved the research team through several possible IT solutions. Among these several dead ends are most of the technologies initially considered key. Leaps of progress are seldom and expensive. Investigation of many unknown paths, of the capabilities of new algorithmic possibilities, is essential to making progress. This highly explorative way of innovation and development was only possible due to the explicit labeling of the project as 'high risk'. It would be all too easy to point to naive planning, overoptimistic estimation of the potential of the technological state of the art, and maybe even to plain bad research. Yet far from being a haphazard experiment, this was a well-planned and executed mid-sized external investment proposal, put through the highest quality assessment criteria of the Netherlands Scientific Organization and subsequently nationally and internationally peer reviewed, and approved with the highest ratings. So according to all the people that should have been able to know, there was nothing wrong with the plan as initially presented. The failure of the Circulation of Knowledge project to execute exactly according to plan is caused by the need of the research team to learn while they are doing. This is the essence of research: experiment and explore, learn, apply.

Yet BAMBOO, DARIAH, CLARIN, and similar projects with big footprints are developing generalized infrastructure for digital humanities research – or so at least they claim. Due to their complex multi-national – in Europe in any case – nature, these projects are planned and executed according to strictly defined project management frameworks aimed at controlling and monitoring large scale projects such as PRINCE2 and PMBOK. These management frameworks are in essence aimed at keeping the risk of project failure in check by careful and meticulous project phasing, planning and monitoring of execution. A so-called waterfall type of software architecture and development is the natural outcome of PRINCE2 project management: a project is divided into

clear and separate planning and control phases: requirements gathering, design, implementation, testing, deployment and maintenance. Each phase includes defined and precise documentation objectives. Given a broad coverage and precise investigation of the user requirements, software development is started with a functional design. This is followed by a technical specification which is a detailed blueprint that software developers can implement into code. The code delivered answers exactly to the requirements, in theory (Baars 2006). Although the waterfall model may work perfectly in cases where it is known from experience quite precisely what the real user needs are, and in cases where technologies and approaches are well tried, it breaks down in research environments. Waterfall models were designed for risk management and reliable execution of fixed plans, not for change management and support. Consequently they have severe difficulty in providing for changing insights, perspectives, requirements and aims, without running into major time or budget overspending. However, change is exactly what tends to happen in exploratory innovative research and development: each single step leads to new insights and thus to changing requirements.

Large scale infrastructure projects are governed through just these sorts of extensive planning, monitoring and control schemes. Anyone having experienced up close the administrative papermill of a Seventh Framework Programme project knows that monitoring and accountability are big time consumers. Accordingly DARIAH, for example, was in a preparatory phase for 30 months (DARIAH 2011). This phase was only concerned with organizing the partner consortium and planning. The fact that a development program needs more than two years on the drawing board makes one wonder how incredibly complicated it must be, to the point that any planning must be symbolic rather than effective. Not only are research and development targets a moving and shape-shifting set of humanities data concepts and workflows, technical platforms and languages change and shift under the feet of development programs too. Thirty months of design and planning is not just long in terms of the average humanities research project, it is an eon in IT time. While DARIAH was in its preparatory phase, in April 2010 the iPad was introduced. Has DARIAH's digital infrastructure design been updated to foresee the potential use of mobile and tablet applications in humanities? There will be new technologies introduced that are relevant to humanities during the execution phase of large infrastructure projects (cf. for instance <http://www.textal.org/>). And if it is to be a generalizable infrastructure catering to all of the humanities, DARIAH should support the development, deployment and use of these technologies that have not even been invented yet.

Essentially, we cannot know what we are supposed to be planning; this makes the sheer complexity of planning, control, design, and platform support within such a large organizational structure even more worrisome, and must have huge consequences for the quality of functional and technical design of

any resulting digital infrastructure. Innovative computationally-aimed research and development in the humanities will rely in every single case on highly specific and constantly changing algorithms and by consequence perpetually shifting implementation technologies. How does one design and plan a general unified infrastructure for a target that is all over the place? A 'one size fits all' approach would be a disastrous underestimation of the specific needs of humanities research. The essence of humanities research is in its diverse, heterogeneous and ephemeral nature. But the need of big infrastructures to be designed well before implementation and use means that every single design decision closes the resulting system to some category of users. Anything specific can potentially break the data model and the defined workflow. Of course big infrastructure architects also know this; to compensate, they look for possibilities to make more abstract models and malleable workflows. But abstract away far enough and the model becomes rather meaningless. If the design recommendation of big infrastructure is that we will use XML, what was the use of all that time and effort? Big infrastructure design seems to routinely underestimate this problem. To give but one simple example: the token seems such a logical, granular, easy to automate conceptual catch-all for text models. Can there possibly be a more generalizable model for text than as a series of tokens? You may think so, until someone asks how you handle languages that do not use spaces, or how you account for prefixes and suffixes, let alone ambiguous functions. Regarding text as a series of tokens denies in any case several other essential humanistic aspects of text: its typography, its materiality, its referentiality to name but a few. At second glance, from the humanities research perspective, there is actually pretty little generalizable about tokenization (Van Dalen-Oskam 2012, 11).

This is the central paradox for big infrastructure design: the very wish to cater to everyone pushes the designers toward generalization, and thus necessarily away from delivering data models specific enough to be useful to anyone.

The Standard Reflex

Since my entry in the field of digital humanities in 2000 I have sat through countless digital humanities project planning meetings. I dare posit that inevitably, at some point during the course of discussion in any such meeting – usually around the time when the subject of a data model is raised – one of the participants will sit up, clear his or her throat and proclaim that we should standardize our terminology; or database fields, or categories, ontology concepts, registry entries, or whatever suitable category of objects that will lend itself superficially to standardization. Standardize, and all the problems that stand in the way of generalized digital tools and infrastructure will disappear.

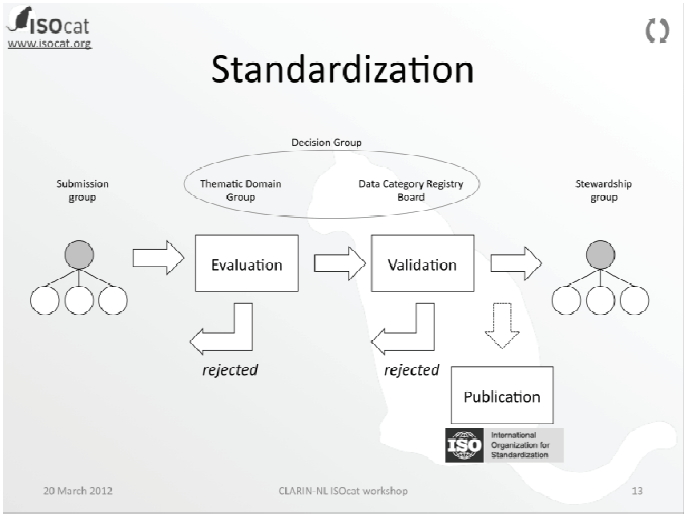
If only everybody would wear size 9 shoes, wouldn't that be a blessing for the shoemaking industry? Standardization seems to be the magic bullet that

many scholars want to throw at the problems that heterogeneous data and specific requirements pose to design, implementation and use of digital infrastructure. But the same paradox of generalization also applies to standards: the exact purpose and need of explorative research is to go beyond what is within the standard. So if a generalized digital infrastructure is to serve any meaningful humanities research it cannot be entirely governed by standards. Declaring that "we will use a standard" is to declare what exactly? That any data or tool not matching that standard will be rejected on these infrastructures? If so, these are going to be mightily empty infrastructures. If an infrastructure really wants to cater to the needs of its users, the only realistic options are to either make room for diversity within a standard – this is basically what the TEI has done by allowing arbitrary definitions to be appended via customization to a de facto standard for text encoding – or to make room for any number of standards that may be used for the same purpose on an infrastructure. The latter is what CLARIN proposes through its ISOCat system. However, both these solutions are counterproductive for generalization. The TEI for example – carefully calling its definitions both a standard and a 'set of guidelines' (cf. <http://www.tei-c.org/>) – is arguably the most successful standard in its subfield of digital humanities, if not in the whole of digital humanities. But the catch-all strategy of the TEI means that it caters for the needs of specific text encoding by explicitly allowing any encoding. That in any case defies the purpose of generalizability of the standard. To quote from the final report of the MONK project:

> One of the declared goals of the Text Encoding Initiative has been to create digitally encoded texts that are 'machine-actionable' in the sense of allowing a machine to process the differences that human readers negotiate effortlessly in moving from a paragraph, stanza, scene etc. in one book to a similar instance in another. American university libraries have developed a six-level hierarchy of encoding texts that is theoretically interoperable, but as we discovered very early in MONK, in practice these texts do not actually interoperate. Encoding projects at Virginia, Michigan, North Carolina, and Indiana certainly share family resemblances, but it is also obvious that in the design of these projects local preferences or convenience always took precedence over ensuring that 'my texts' will play nicely with 'your texts'. And aside from simple interoperability, there is even less affordance for extensibility: none of the archives seriously considered the possibility that some third party might want to tokenize or linguistically annotate their texts (Monk 2009, 4-5).

A similar effect undermines the ISOcat system devised by CLARIN. The ISOcat system describes itself as a data category registry, "Defining widely accepted linguistic concepts" (<http://www.isocat.org/>). Once accepted, the standard described within ISOcat will be usable on CLARIN infrastructure; the intent is to gather standards until a domain (in this case the linguistics domain) is sufficiently covered. Governance of standard acceptance is through a Decision Group that is composed of a Thematic Domain Group and the Data Category Registry Board (cf. Fig. 2).

Figure 2: ISOcat Standards Governance Flowchart



Interestingly, the decision making flow chart has two 'rejected' states, but no 'accepted'. However, just as with bloating of a single standard, generalizability is not well supported by ballot.

Apart from being quite fictitious instruments of generalization, what purpose may be found in standardization for boundary-pushing research? When they become a goal in themselves rather than a means, standards run the risk of impeding rather than leveraging innovation and research. There can be no absolutism in standards conformance if we value open research practices. When proclaiming yet another standard as a solution we should never forget that there is actually no such thing as a true standard. The metric system is not absolute even if it looks more logical than having three feet in a yard. Power sockets keep the adapter manufacturers alive. And a US size 9 is called 40 in Europe. And even if standards are not absolute, or maybe because of that, we have more than enough standards. Just for the purpose of creating and maintaining metadata in the cultural heritage sector the count is at least 150 (Riley 2009). The number would multiply manifold if we need to draw in also all IT standards that are available.

Thus, the purpose of standards seems to defeat itself, and a 'standards based generalized digital infrastructure' that would cater to all standards … is not that just the Internet then? When making claims about generalized infrastructure, are we claiming that it be inclusive or exclusive? The very governance of CLARIN, for instance, unfortunately seems to suggest the latter. But how can we then seriously maintain that we are working toward open and generalized infrastructures? Furthermore I doubt if digital humanities in itself represents a

large enough user base to push any standard for IT infrastructure. The only standard that comes close would indeed be TEI. But that of course is an encoding standard, not an infrastructure standard. In general, IT standards are set by industry and adopted by the humanities research sector, rather than the other way round. Thus, we see that the convergent XML format for textual content in industry is ePub rather than TEI.

Embracing Change

Large all-purpose digital infrastructures meant to serve all users are dependent on generalization through standards in order to sustain themselves and remain under institutional control. But as we have seen, innovative tools and methodology sprout in a realm of relative uncertainty of technology and requirements. This is precisely the opposite of what big infrastructures are likely to support. The quick technology shifts and development of thinking that research innovation requires cannot be supported through these large unified infrastructures. In this sense, large standards-based infrastructures will necessarily be intellectually prohibitive places: they do not allow the conception of new ideas, new approaches, or new models, as these do not fit the required infrastructural mold. What fits in the box, fits in the box, but out of the box thinking is an unsupported feature.

The opposition between the wish for efficient development and support of large controlled infrastructures for sharing unified resources on the one hand, and the need for quick and flexible short-lived solutions developed along evolutionary lines on the other hand, are by no means unique to digital humanities. The problem of tension between innovation and consolidation in fact has sparked entire theories of management (Poppendieck 2009) and software development processes (Beck 1999). As large industry IT development projects ran aground more and more in the late nineties and early years of the new millennium, it became clear that something was wrong with the waterfall type of monitoring, control and risk management in many cases of software development. A group of software developers decided to adopt a different approach to software engineering and infrastructure development which resulted in the so called Agile Process (Fowler 2012) methodology. As with any new movement, various sub-factions arose, some more dogmatic than others. Whichever variant is chosen, it is my experience that if the core principles are applied, they lead to working, efficiently produced, usable software. It is hard to prioritize agile principles, and in fact I would advise against doing so in the urge to cut corners on development process, but if pressed to choose the most important ones I would say ‘value humans and interaction over planning and documentation’, ‘realize the simplest thing that could possibly work’, and most of all: ‘value responding to change over following a plan’, which usually gets shortened to ‘embrace change’.

It is especially that last principle that, in my experience, supports research driven software development very well. Agile software development works in short bursts of creativity, called iterations or sprints, which can be as short as a week, or even less, but never more than 3 weeks. A sprint begins with a discussion between researcher and developers on what needs to be developed; it ends with the evaluation of the results by the same researcher and developers. The next iteration is planned as an answer to the changes in thinking that the experience provoked in both the researcher and the developers. In this way the actual tool or software evolves ever more into what the particular researcher actually needs, and not what some design committee thinks might be needed by all researchers.

The nature of lean or agile approaches to tool and infrastructure development also ensures that any infrastructure that is delivered will be as lightweight as possible, according to the principle to implement only the simplest thing that could possibly work. The need to respond to change also keeps the result lightweight; the heavier the technical footprint of a solution – that is, the more software and hardware needed – the harder it is to swap out parts of the solution for more suitable technologies. The same principles lend themselves to code reuse, as it is far simpler to reuse existing solutions and to adapt those to changing needs than to reinvent solutions.

It is these lightweight agile approaches that underpin the successes of projects like Huygens ING's 'Interedition', The Center for History and New Media's 'One Week, One Tool', MITH's 'XML Barn-Raising', as surveyed by Doug Reside. Reside concludes that by spending time in rapid prototyping rather than standards discussions for hypothetical use cases, by gathering scholars who are also coders (rather than those with only one skill or the other), these work sprints quickly determine the real problems facing infrastructure development and often make significant headway towards solving them. Indeed these lightweight development projects have delivered more new tools – such as Anthologize (<http://anthologize.org/>), CollateX (<http://collatex.sourceforge.net/>), ANGLES, Stexaminer (<https://github.com/tla/stemmatology>) and more – and progressed the development of more existing technologies for projects – such as TextGrid (<http://www.textgrid.de/>) and Juxta (<http://www.juxtasoftware.org/>) – than any of the big infrastructure projects. This leads me to share Reside's conclusion:

> I am convinced that code camps are a better investment than the large digital infrastructure projects and, with some improvements, have the potential to revolutionize scholarly and library technology development ecosystem. They are, after all, many times cheaper to run. Funding agencies might consider whether what has been shown to be possible with tens of thousands of dollars or euros should be funded with millions (Reside 2012).

Apart from being cost efficient, highly focused, and congruent with the nature of humanities research practice, I think another important characteristic of

lightweight approaches to tool development is that they are – or at least can be – more self-sustaining through being fully community-rooted. This is not to say that there is no cost. It rather means that the tools so far produced under lightweight development processes are all open source projects maintained by a non-institutionalized community of researcher-developers, including even primary users of the tools. This community aspect ensures that the components and software libraries produced do not exclusively cater to the needs of the owning institution. The shared stake that a development community has in a certain tool is one of the most important properties that drives the focus on purpose and application; moreover, it ensure an incentive to maintain and support development as long as the tool serves a purpose. The possibility is very real, of course, that maintenance and support will dwindle and an application may die if primary users lose the need for a tool. I wonder though if this is a problem: unused tools should die rather than become a burden on already scarce development and support capacity. In this sense lightweight approaches are also self-validating.

From Encoding to Modeling, the Changing Research Practice

Part of my argument against large research infrastructures developed in a top-down fashion is that they bear a heavy inflexible footprint in their technology and standards, and are therefore ill-suited to supporting the heterogeneous and ephemeral characteristics of humanities data and research. One could argue that this is not a problem as long as large digital infrastructures are aimed only at hosting and safeguarding research data. Given machine-negotiable access services to that data, then any tool of any kind might be applied. Tools that are relatively easy to generalize (concordancing services for instance) could be maintained more stably on such an infrastructure too. Yet this would still allow for 'agile development space' to add and use tools on less institutionalized infrastructure.

However, digital tools in the humanities are becoming an expression of research itself. Until recently one could probably maintain that, within the digital textual scholarship community, the foremost technology used to capture text structure semantics was through encoding (markup), preferably in TEI-XML format. The encoded digital text was therefore the most prominent and useful digital expression or result that could be sustained. But with more and more tools appearing that process, interpret, and visualize text, this is changing. A tool like CollateX is based on a carefully analyzed heuristic model of the scholarly text collation process which is dissected into discrete steps, each of which is modeled into code (Haentjens Dekker 2011). CollateX in itself thus represents a heuristic model for text variation analysis. But there are more ways to approach text variation analysis. The nCritic (<https://github.com/tla/

ncritic>) automatic collator, for instance, uses a slightly different model to collate texts. What this shows, essentially, is that tool building is not a mere research-independent act to enable data processing. Rather, it is the act of modeling humanities data and heuristics as an intrinsic aspect of research. Tool and software development thus represent in part the capture and expression of interpretations about structure and properties of data, as well as interactions with that data. This type of lightweight, highly specific, research-driven tool development is therefore reminiscent of – and possibly even a reification of – the ideas on modeling put forward by people such as Orlandi and McCarthy (Rockwell 2012). It would plainly defy the purpose of large scale digital infrastructures for humanities if they were not also to host and sustain such tools. It is akin to saying "we will maintain the text, but not the book".

In the case of digital textual scholarship, the encoded text file will not be the start and end of humanities text data capture and storage. As Peter Boot and I argue (Van Zundert 2011), future digital scholarly editions will look less like self-contained digitized books. Digital editions will instead be interfaces for engaging with digital text, created by combining services that process text and related content from sources distributed on the Internet. For instance, a simple digital edition might be constructed from a few independently hosted services. One fetches the facsimile images from a server at the British Library, the second fetches the transcribed text from a server of Sankt Gallen University. A third service combines the results in a pageable representation of text and images per page. Such a 'Web 2.0' web services based edition can be fitted with a tool to register user comments that are stored on yet another server offering an annotation service. The edition might also be fitted with a tool to suggest transcription corrections or even to directly edit the transcription. These Service Oriented Architecture – or SOA for short – possibilities that are now becoming the bread and butter of Internet technology have the potential to transform digital scholarly editions from what are essentially reading interfaces into virtual research environments based on distributed technology. Any large digital research infrastructure claiming to support digital scholarship should be able to support and maintain distributed 'Web 2.0' editions. Moreover, it should be able to capture the potentially perpetual modifications and additions that are made to the edition as it is used and annotated by researchers and other users.

These ideas on text as data that can be processed, text that can be transformed through services, and scholarly editions as working environments for texts in indefinite progress – ideas that are, of course, by no means new (cf. for example Buzzetti (2006), Robinson (2005) – are technological avenues for text that we need to explore – both to discover the potential representations and uses of text in the virtual environment, and also to add to our formalized understanding of the properties of text. It is this kind of fluidity and virtuality of text as data that should be supported by any generalized humanities infrastructure. But given that interfaces and services for these ideas are still very early imma-

ture technologies in a wide variety of experimental setups, support indeed seems a daunting task for a current state-of-the-art standards-based infrastructure. In fact, even limiting support for the ideas for representing text and interfaces to what is possible with current mainstream internet technology, along the lines that Pierazzo (Pierazzo 2011) or Rosselli Del Turco (Del Turco 2012) point out – that is, even if we refrain from trying to extend our models for text representation and interaction beyond what we can practically achieve today – the task is beyond any currently feasible large-scale generalized digital infrastructure for the humanities.

In other words, transformed by digital technology, text and digital editions – digital humanities data in general, as a matter of fact – become fluid, 'living', reaching a state wherein they are perpetually in a digital information lifecycle. Providing a generalized digital infrastructure for the humanities for such volatile research data – and indeed for such volatile data modeling – is a huge development and maintenance challenge. And even if a generalized model for volatile data could be delivered right now, there would be the additional complication of providing various specific user interfaces for it. For again one size will not fit all here as each researcher will have individual needs and requirements to put forward – not on the basis of some narrow-minded aesthetics but on the basis of specific research questions. What is needed is not just the support to store text representation according to encoding models. To really support textual scholarship and research also requires the ability to model text and text interaction into dynamic data models and algorithmic models, to put those models to work as research-specific services on the infrastructure in question, and to support change in the models. That is rather different from implementing a current state-of-the-art markup approach for digital editions on an institutionalized grid.

Sustainability

One of the oft-stated purposes of large-scale digital humanities infrastructure is to preserve and safeguard the tools and data pertaining to humanities research that are produced, and to share the technical and organizational burden for that task. Digital libraries would feel they have a stake in here, and a role to play (Van Zundert 2011). The assumption that digital libraries and research institutions seem to make is that institutional collaboration on the erection and upkeep of such an infrastructure will be a good warrant that the infrastructure will be there indefinitely. In part this is true, for indeed it is unlikely that libraries, universities, and research institutions will vanish any time soon. And indeed spreading or sharing the burden for digital archival infrastructure is sensible as the risks of administering data, tools, and services at just one institute – or even worse on one machine – are severe, of course. Monolithic systems are single points of failures. But what is more, the burden of sustaining local monoliths

may even pose a threat to the institution of overburdened resources. The institution needs, at the very least, redundant storage and server capacity to ward off the greatest risks of technical failure. But then, one needs safe remote copies too. System, server, network and application support are needed for the maintenance of all machines and all software on them. And as tools and data come in all varieties, the list of scarce IT knowledge of software and standards that needs to be available 'in-house' explodes within a short space of time, and starts eating away valuable human and non-human resources. Thus sustainability through institutional maintenance of digital infrastructure is a large and potentially explosive burden.

However, institutional stability and integrity are not the only aspects we should take into account. In the preceding section I suggest that there is unlikely to be a very clear cut off in the future between 'dead archival data' or 'finalized editions' on the one hand, and 'living data' and/or 'works in progress' in virtual research environments. The distinction between those concepts and – crucially – between the technical implementation of the concepts may become permeable, may even fully disappear. The classic information lifecycle (Boonstra 2004) might soon no longer have a neat off ramp where results are ported to a stable archival silo. Rather in contrast, the information in a majority of cases will keep spinning around within that very lifecycle. The current push of Internet technology towards Service Oriented Architecture (SOA) may very well reinforce this change. Known by their buzzword 'The Cloud', service oriented architectures are nothing more or less than data-publishing software processes running as services on the same Internet we already know. Until now the Internet has been mainly used to 'display data' for human consumption; the innovation of the last few years is the exposure of data in machine negotiable form and the ability to publish services, i.e. software that can process data exposed on the Internet. To put it more simply perhaps: where formerly you would necessarily have data and software working together on a single computer, processes and data may in the future be stored and executed on any computer connected to the Internet. We may at some point not even know which data is on what computers, and any service may be executed by various machines. A service may even be executed in part by one machine and in part by another. This may sound like dangerous chaos to a librarian, but from an IT perspective it makes perfect sense if sustainability for tools and data is key. When data and services can be replicated automatically on any node of a large network of computers, the chances of date loss due to a single point of failure are virtually zero. This strategy of sustaining data and tools is based on keeping redundant copies and ensuring that services and data are stored in several redundant locations. In fact, this is exactly what LOCKSS (<http://lockss.stanford.edu/>, Rosenthal 2011) is doing. The approach is sound and is based in essence on the same principles that make guerilla warfare the hardest to

tackle, made the Victory Ships an overwhelming success, and make computer viruses and music piracy so hard to eradicate.

Such completely self-replicating and load-balancing networks of services and data may render digital information archival in the classic 'data storage' sense obsolete. Service Oriented Architecture, Cloud Computing, Distributed Computing are all technologies that push in the direction of an Internet of services where institutions, industry and individuals share computing and storage capacity on an on-demand basis. These are highly cost-efficient ways to maintain and distribute network capacity. These are also technologies where sustainability is a part of the running system as it were, rather than a task of separate storage silos and institutions. Such technologies are more in tune with a research perspective wherein we work with living and fluid data objects in continuous research life cycles. These are also the sorts of technologies that enable us to advance our data modeling and analysis.

Modeling Workflows with Microservices

Over the last few years, the Interedition project has been experimenting with service-oriented models as described here. The project has delivered several web-enabled workflows based on the idea of microservices. Microservices are specialized instances of web services that take their data in JSON format over a RESTlike protocol, and return the data processed in a way that is meaningful within a certain research workflow. This is arguably the most lightweight implementation that is possible for any web service. JSON (<http://www.json.org/>) is the simplest data structure format in existence, yet is expressive enough to encapsulate almost any higher-order data structure such as XML or object models. A more basic client-server communication protocol than REST (<http://en.wikipedia.org/wiki/Representational_state_transfer>) is hard to imagine.

Individual microservices represent discrete steps of scholarly processes that can be individually automated (Haentjens Dekker 2011). Such microservices can be chained together into 'pipes' to construct larger workflows for scholarly purposes. Prototype webservice-based workflows that Interedition delivered constituted amongst others a full textual collation workflow (<http://collatex.sourceforge.net/)[2], a fully-implemented prototype OAC annotation service

[2] The current demo package for CollateX is offered not as a chain of microservices online, but as a Java Webstart application able to run on any local computer. This is due to response time constraints by Google App Engine, to which the web service chain could not comply under prototype conditions. Interedition promised proof-of-concept prototypes only, however, demand for a production level implementation was concrete enough that the decision was made to publish the CollateX tool in this form too. The proof-of-concept mi-

(<http://demo.interedition.eu/raxld/>), an ngram extractor (<http://www.interedition.eu/wiki/index.php/Ngram>), and several more. These microservices run on no other infrastructure than the current Web and need no other support than standard off-the-shelf open source server platforms. In other words, apart from the Internet they need no additional infrastructure.

These super-lightweight humanities research microservices have considerable advantages over any integrated or purpose-built large digital infrastructure. They can be deployed on any mainstream server platform – or, for that matter, 'in the cloud'. They are open source and collaboratively built. This ensures shared knowledge about their inner architecture and implementation, and gives other developers the ability not only to maintain them but also to reuse them. They are implementation-agnostic: it does not matter if they are implemented in Java, Ruby, Perl, R, or any other language; as long as they serve JSON over a REST protocol they can be incorporated into a scholarly workflow. This implies that anybody who can code can contribute, regardless of their preferred implementation language or architecture. This makes the whole philosophy behind microservices open, inclusive, lightweight, and sustainable.

The Broken Business Model

True modeling of humanities research data and heuristics through code and through the data itself actually requires very little in the way of specialized large-scale humanities research infrastructure. What those requirements amount to is the availability of enough mainstream network, storage, and computing capacity, which is to say: 'enough Internet'. There is little rationale to put more effort than is needed to organize that on the technical infrastructure level. The most useful thing big infrastructure programs could do, therefore, is to make a considerable investment in an academic cloud, allowing access, storage and execution to any member of the academic endeavor. We all have email, why not let us all have access to an academic computing cloud? The technologies for basic functions such as single sign-on, governance, persistent addressing, service brokerages and orchestration are all readily available through industry. Even storage – which may seem like a problem because we are ever storing more data – will rapidly become trivial as long as Moore's Law (<http://en.wikipedia.org/wiki/Moore%27s_law>) still holds.

Of course there are some darker wisps of cloud in this blue sky scenario. The foremost in my view is that the autonomous cloud approach that potentially lies in the future can make accounting rather opaque for institutions who wish to know exactly which data and services they are hosting, or whether their

croservice chain is also on line at http://interedition-tools.appspot.com/ but is likely to show response time errors.

effort and financial investment are indeed completely proportional to their use of cloud facilities. To understand this we must return for a moment to the idea of the digital scholarly edition. Now that the edition might well be a perpetually active research environment into which tools can be plugged in and out, where do the responsibilities for maintaining such an edition lie? How do we find the point at which the developing research institution transfers maintenance to a preservation-aimed institution, such as a digital library? This broken business model is rather more of a problem to be solved for the future than the infrastructure issue.

This problem, complex as it is, will be further complicated when we move from 'Web 2.0' digital humanities to 'beyond web' technology. The 'Web 2.0' digital edition as a virtual research environment based on orchestrated distributed microservices may sound like a farfetched idea – although it is not so outlandish, inasmuch as Huygens ING is implementing the model for its eLaborate digital edition framework – but the future is even more spectacular. IBM predicts (<http://www.research.ibm.com/autonomic/>) that autonomous systems and autonomous computing are the next step in general computing and digital networks. Autonomous computing is the idea that code and data can exist on a digital network such as the Internet independently of any local server. As things stand today, one needs to deploy (i.e. install) services on a specific server, which means that the service will run on that specific machine, indefinitely in principle. Autonomous computing 'liberates' an algorithm – and any associated data – from its confines on a local server, as a bird from its cage. The application can travel to any other server in the network based on where the network needs its services the most. Hence imagine: this is not your text file moving from one computer to another by mail, but it's your text processor moving – while running. The utility of this is that software can travel to where it is needed and can be executed in the proper context. The ability to move active code and data as singular digital object through a network opens up the possibility to think of – for instance – a digital scholarly edition as 'living data' in the sense that it is fluid and editable anywhere, encapsulated by all the code and interface that is needed to edit, authorize modifications to, re-version, or annotate the data, or perform any other imaginable task. We might then begin to think of the book as an active object with inbuilt behavior. As a researcher of literary texts and a developer of digital literary curation and analysis tools, pondering the potential of such a transformation of text into the digital is far more exciting to me than the properties of any current digital infrastructure technology.

Moving Forward

The inevitable conclusion from all of the above must be that, at least as far as digital humanities research is concerned, there is little benefit to be expected

from the current large infrastructure projects. Their all-purpose nature enforces a generalized strategy aimed at the establishment of standards which is at odds with innovative, explorative research. Being standards-driven, institutionally bound, and at worst enforcing specific implementations, they are platforms of exclusiveness. But the field of digital humanities is still maturing; it is embedded in and it supports a research domain that is based on heterogeneous data and divergent research questions. Nascent and ever evolving, digital humanities needs open and inclusive platforms. It is the lightweight, agility-based, low cost projects that have demonstrably delivered more useful tools and models for building the tools than any of the big infrastructure projects have. In part this is probably because tool building in the digital humanities is still in the process of maturing as well, and is still exploring the development processes that are a good match for humanities research it supports. However, as we have seen, these agile processes combined with a webservice-based approach seem to foster evolutionary development towards useful and usable tools that are maintainable and sustainable. Moreover, these approaches are far better suited to follow the changing and shifting properties of the larger IT context.

The wish for digital infrastructure in itself is well motivated, for of course digital humanities needs a sand box and building ground. However, these infrastructures should indeed be the simplest thing that could possibly work. That infrastructure is actually already out there and is called the Internet. If institutions are truly committed to supporting tool development and data modeling, then they should focus on knowledge exchange between digital humanities developers and researchers, on allowing them to work together in code challenges and work sprints, on investment in digital humanities curricula, and in academic credit for the results of digital humanities activities such as tools and models.

Coding and modeling are more than just collateral of the academic activities within digital humanities; they are central to the whole enterprise. If we shift our central focus here, and take the infrastructure itself as less central, we will create the right context for truly groundbreaking engagement with humanities research data in virtual environments. What we do not need is precisely the bulky concrete highways; we can make do with the landscape that is already taking shape out there. Some bricks, mortar, shovels and gravel would be nice though, as well as a manual on how to use them.

References

Baars, Wouter, Henk Harmsen, Rutger Kramer, Laurents Sesink, and Joris van Zundert. 2006. *Handboek project management DANS, met speciale aandacht voor de werkwijze bij projecten voor software ontwikkeling*. Den Haag: DANS.

Beaulieu, Anne, Karina van Dalen-Oskam, and Joris van Zundert. 2012. Between tradition and web 2.0: eLaborate as social experiment in humanities scholarship.

In *Social Software and the Evolution of User Expertise*, ed. Tatjana Takseva. Hershey, PA: IGI Global.

Beck, Kent. 1999. *Extreme Programming Explained: Embrace Change*. Addison-Wesley Professional.

Boonstra, Onno, Leen Breure, and Peter Doorn. 2004. *Past, Present and Future of Historical Information Science.* Amsterdam.

Boot, P. 2011. Towards a Genre Analysis of Online Book Discussion: socializing, participation and publication in the Dutch booksphere. In *Selected Papers of Internet Research. IR* 12.

Buzzetti, Dino, and Jerome McGann. 2006. Critical Editing in a Digital HorizonIn *Electronic Textual Editing*, ed. Lou Burnard, 53-73. New York: MLA.

Crane, Gregory. 2006. What Do You Do with a Million Books? *D-Lib Magazine* 12 (3), <http://www.dlib.org/dlib/march06/crane/03crane.html>.

Clariah. 2012. Common Lab Research Infrastructure for the Arts and Humanities. CL@RIAH, <http://www.clariah.nl/?q=node/4> (Accessed April 11, 2012).

DARIAH. 2011. "History." DARIAH-EU, <http://www.dariah.eu/index.php?option=com_content&view=article&id=7&Itemid=119> (Accessed April 11, 2012).

Del Turco, Roberto Rosselli. 2011. After the editing is done: Designing a Graphic User Interface for digital editions. *Digital Medievalist* 7, <http://www.digitalmedievalist.org/journal/7/rosselliDelTurco/> (Accessed April 11, 2012).

Fish, Stanley. 2012. The Old Order Changeth. *New York Times*, December 26, 2012, <http://opinionator.blogs.nytimes.com/2011/12/26/the-old-order-changeth/> (Accessed April 11, 2012).

Fowler, Martin. 2012. Agile Software Development. Fowler, <http://martinfowler.com/agile.html> (Accessed April 11, 2012).

Gray, Catherine. *If you build it, will they come? How researchers perceive and use web 2.0. A Research Information Network report.* London: RIN.

Haentjens Dekker, Ronald, and Gregor Middel. 2011. Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements. Paper presented at Supporting Digital Humanities 2011, Copenhagen, Denmark, November 17-18, 2011.

Interedition. Digitally Sustaining our Textual Heritage. Huygens ING. <http://www.interedition.eu/?page_id=84> (Accessed April 11 2012).

KB. 2010. Koninklijke Bibliotheek and Google sign book digitisation agreement. Netherlands Royal Library, <http://www.kb.nl/nieuws/2010/google-en.html> (Accessed April 11, 2012).

Pierazzo, Elena. 2011. A rationale of digital documentary editions. *Literary and Linguist Computing* 26 (4): 463-77. doi: 10.1093/llc/fqr033.

Poppendieck, Mary, and Tom Poppendieck. 2009. *Leading Lean Software Development: Results Are not the Point*. Addison-Wesley Professional.

Prescott, Andrew. 2010. To Code or Not to Code? Digital Riffs, blog posted April 7, 2010, <http://digitalriffs.blogspot.com/2012/04/to-code-or-not-to-code.html> (Accessed April 11, 2012).

Ramsay, Stephen. 2011. Who's In and Who's Out. Stephen Ramsay, <http://lenz.unl.edu/papers/2011/01/08/whos-in-and-whos-out.html> (Accessed April 11, 2012).

Reside, Doug. 2012. Smaller is Smarter. Paper presented at the Interedition Symposium Scholarly Digital Editions, Tools and Infrastructure. The Hague, The Netherlands, March 19-20, 2012.

Riley, Jenn. 2009. Seeing Standards: A Visualization of the Metadata Universe. Indiana University Libraries, <http://www.dlib.indiana.edu/~jenlrile/metadatamap/> (Accessed April 11, 2012).

Robinson, Peter. 2005. Current issues in making digital editions of medieval texts – or, do electronic scholarly editions have a future? *Digital Medievalist* 1, <http://www.digitalmedievalist.org/journal/1.1/robinson/>.

Rockwell, Geoffrey. 2012. Digital Humanities in Italy: Tito Orlandi. Theoreti.ca, blog posted April 14, 2012, <http://www.theoreti.ca/?p=4333> (Accessed April 14, 2012).

Rosenthal, David. 2011. How Few Copies? Rosenthal, blog posted March 15, 2011, <http://blog.dshr.org/2011/03/how-few-copies.html> (Accessed April 11, 2012).

Stacey, Ralph D. Stacey. 1996. *Complexity and creativity in organizations*. San Francisco: Berret – Koehler Publishers..

Unsworth, John, and Martin Mueller. 2009. *The MONK Project Final Report.* Indiana University, the University of North Carolina at Chapel Hill, the University of Virginia, and Martin Mueller at Northwestern University: Monk.

Van Dalen-Oskam, Karina. 2012. Names in novels: an experiment in computational stylistics. Literary and Linguistic Computing Advance Access. Oxford: Oxford University Press. doi:10.1093/llc/fqs007Van Zundert, J., and P. Boot. 2011. The Digital Edition 2.0 and the Digital Library: Services, not Resources. *Digitale Edition und Forschungsbibliothek* 44: 141-52.

Van Zundert, J., Smiljana Antonijevic, Anne Beaulieu, Douwe Zeldenrust, Karina van Dalen-Oskam, and Tara L. Andrews. 2012. Cultures of Formalization Towards an encounter between humanities and computing. In *Understanding Digital Humanities*, ed. David Berry, 279-94. London: Palgrave Macmillan.

Welshons, Marlo, ed. 2006. *Our Cultural Commonwealth, The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. Urbana-Champaign: ACLS/University of Illinois.