Sustainability Strategies for Digital Humanities Systems

Claes Neuefeind[1], Philip Schildkamp[2], Brigitte Mathiak[2], Unmil Karadkar[3], Johannes Stigler[3], Elisabeth Steiner[3], Gunter Vasold[3], Fabio Tosques[3], Arianna Ciula[4], Brian Maher[4], Greg Newton[5], Stewart Arneil[5], Martin Holmes[5]

[1] Cologne Center for eHumanities, University of Cologne, Germany
[2] Data Center for the Humanities, University of Cologne, Germany
[3] Centre for Information Modelling, University of Graz, Austria
[4] King's Digital Lab, King's College London, United Kingdom
[5] Humanities Computing and Media Centre, University of Victoria, Canada

Now that the Digital Humanities (DH) are becoming a well-established research field, producing seminal publications in print as well as digital formats, the time for consolidation has come. It is noteworthy that digital tools and methods from the pioneering days of the DH are degrading and some have already vanished. Therefore, it is urgent to take action and to prevent further losses. While the necessity of high quality research data management (RDM) is encouraged or even required by funding agencies and there is an increasing awareness for long-term archiving (LTA), when it comes to primary research data, the fact that the DH exhibit a structural deficit regarding maintaining and preserving research software is at the least underestimated.

In this panel, we will focus on infrastructure and institutional support. Beginning with an overview of existing strategies from the DH and beyond, we highlight selected strategies to compare how they are implemented at different institutions in terms of infrastructure, expert knowledge and also funding. We also want to evaluate the extent of institutional support that is needed to successfully sustain and archive DH projects and the software they use. We will discuss currently implemented solutions to maintain and preserve research projects and software, all of which approach the outlined problem from a different angle.

1. Sustainability Strategies in DH and beyond (Brigitte Mathiak, Data Center for the Humanities, University of Cologne)

Sustainability of research software is an important issue for the DH. In our investigation of the "Digital Scholarly Editions" online catalogue, we compared the time stamps of the last seen version on the Internet Archive with the first seen version (Schildkamp & Mathiak, 2019). We discovered that of 466 digital editions, 376 had disappeared (cf. fig. 1). The average life time is 8.5 years, while the half life time is about 6 years. We expect that other DH projects exhibit similar trends. The reasons for the disappearance of these valuable research resources are manifold: diminishing funding, lack of institutional support and, over time, lack of personnel support as researchers switch career paths or research directions. The "Digital Dark Age" (Whitt, 2017) affects not only our digital cultural heritage, but also the born digital outcomes of scholarly labor.

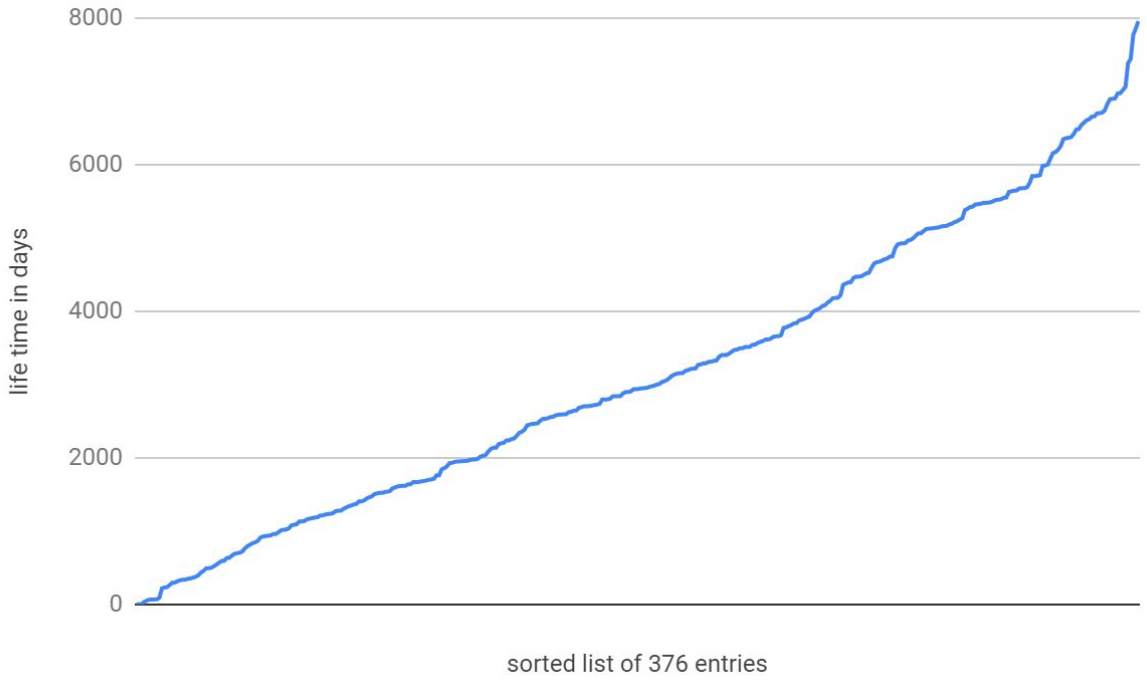

Figure 1: Life time of digital scholarly editions

The problem of sustainability is neither unknown, nor without solutions. Several different models have been explored within the DH community. These include the development of centers such as CHNM, consortia such as Europeana, Hathitrust, and DPLA, as well as community partnerships such as Samvera (previously Hydra) and Islandora. Individual institutions such as those represented on this panel have taken up responsibility for the resources that were placed in their care. Yet, there is a dazzling variety of strategies, technologies, and policies that have been adopted to improve the elusive sustainability, e.g. code archiving, open source dissemination, duplication, sandboxing, refactoring, unified tech stacks, virtual research environments, virtualization, and use of the Internet Archive. While it is clearly easier to prepare a project for sustainability in the planning stage, advice for enhancing sustainability is divergent, ranging from using simple technology, someone's preferred infrastructure, or particular documentation practices. Many completed projects do not have a sustainability strategy, either because they were too old or too optimistic. What happens to these projects is often determined by funding and institutional support. The luxury version is a complete redesign with all the newest bells and whistles, but there are also cheaper strategies, such as putting the system in a sandbox, or relying on the Internet Archive.

However, the problem of sustainability is not unique to the DH. Basic sciences (biology and physics), atmospheric and space sciences, as well as geosciences are some disciplines that are developing sustainability enhancing mechanisms. In conjunction with funding agencies such as the National Science Foundation, researchers in these disciplines have attempted approaches such as community engagement in software and schema development (Specify), ongoing external funding for maintenance (rather than only for new research), long-term funding arcs (NSF centers), funding agency mandates (contribution of digitized

data to existing repositories), efforts to desilo or integrate resources (iDigBio, iPlant), and institutional support for pre-publication drafts (arXiv). We will explore the breadth of these approaches as well as the expected and actual impact of these strategies on sustainability of products that are critical for scholarship in these disciplines, and connect the dots by drawing parallels to the DH.

2. TOSCA-based Application Management (Claes Neuefeind and Philip Schildkamp, Cologne Center for eHumanities/Data Center for the Humanities, University of Cologne)

The University of Cologne's Data Center for the Humanities (DCH) is obliged to concern itself with the sustainability of all digital artifacts produced during (Digital) Humanities projects, e.g. as run by the Cologne Center for eHumanities (CCeH). And as such, it is not only committed to the long-term preservation of data, but of so-called "living systems" (Sahle & Kronenwett, 2013) as well. With regards to this necessity, the DCH is currently engaged in the DFG-funded project "SustainLife - Sustaining Living Digital Systems in the Humanities" (Neuefeind et al., 2018), conducted in cooperation with the Institute of Architecture of Application Systems (IAAS) of the University of Stuttgart. The project aims at adopting the "Topology and Orchestration Specification for Cloud Applications" (TOSCA) standard (OASIS, 2013 and 2019) to the field of Digital Humanities. Being an industry standard focussed on deployment and maintenance of complex software services, TOSCA allows to model applications as abstract topologies consisting of reusable components, while avoiding any kind of vendor or technology lock-in. Through this meta-modelling of software components, not only can the deployment context be adjusted easily (e.g. deployments geared towards OpenStack can easily be adjusted towards Docker, VMWare vSphere, etc.), but from the reusability of said components, synergetic effects emerge, lessening the overall administrative costs for long-term archiving and deployment of research applications.

In our contribution to the panel, we will present the methodological concept of our approach based on the OpenTOSCA ecosystem (Breitenbücher et al., 2016), an open-source implementation of the TOSCA standard, as well as a distinct set of use case implementations conducted within the SustainLife project. The use cases to be presented will cover some of the typical technology stacks in the DH. Foremost, the (1) EarlyCinema use case stands for one of the most common technology stacks: LAMP (Linux, Apache, MySQL, PHP). Further, the (2) AutoPost and (3) TiwoliJ use cases employ the popular Java framework Spring(Boot) with a MySQL database as persistence layer. Also implemented using Spring(Boot), but persisting data in MongoDB, employing ElasticSearch as indexing service and packing a ReactJS frontend, the (4) VedaWeb use case represents one of the more specialized stacks. And lastly the (5) Musical Competitions Database is the most specialized use case, as it depends on older versions of CouchDB for data persistence and ElasticSearch for indexing persisted data (Neuefeind et al., 2019).

3. GAMS: Geisteswissenschaftliches Asset Management System (Unmil Karadkar, Johannes Stigler, Elisabeth Steiner, Gunter Vasold, and Fabio Tosques, Centre for Information Modelling, University of Graz)

Recognizing the problems inherent in conducting digital humanities research based on stand-alone, custom software, the Centre for Information Modeling has developed, maintained, and enriched GAMS--a modular, standards-based, community-used software--since the early 2000's, gaining over 15 years of experience in sustaining a digital scholarship infrastructure. The GAMS infrastructure is supported by ongoing relationships with researchers, personnel, processes, and certifications that inform a holistic, long-term sustainability philosophy. Thus, GAMS embodies a strategy for digital preservation that has been hardened through software upgrades, continuous use, and external testing. GAMS hosts over 95,000 compound digital objects and supports over 90 digital humanities projects.

- Infrastructure:
  The GAMS software is developed using open software and platform-independent standards. These include FEDORA--a flexible open repository infrastructure, Blazegraph, a standards-based, high-performance graph database, Handle--a persistent identifier service, PostgreSQL, Apache Cocoon, Apache Lucene, Apache Solr, and LORIS IIIF image server. GAMS was initially developed using FEDORA 2.0 and over the years, has been migrated to FEDORA 3.5. The GAMS team has developed OAIS-compliant workflows in order to support long-term preservation. Data stored in GAMS is subject to FAIR data principles.
  Currently, the GAMS team is updating the backend to FEDORA 6.0. This upgrade presents unique challenges as FEDORA has outsourced the notion of content models since version 4.0 and model compliance must now be handled in the application layer. The modularity of the GAMS architecture facilitates such an upgrade as the Java-based Cirilo client supports the management of a legacy layer while migrating to a REST-API-based interface. Cirilo is developed on an open source philosophy and is available for download via GitHub.
  The user interface layer of the GAMS Web interface is based on Web technology standards, such as XML and XSLT that separate structure from content and enable multiple, context-specific renditions of Web-based information.

- Relationships:
  In order to ensure continued relevance, the GAMS team partners with humanities researchers. GAMS receives and stores data in recognized archive-compliant standards such as JPEG2000, TIFF, TEI, and LIDO. In addition to providing interfaces for tasks such as the upload, management, description, presentation, and dissemination of digital objects, the team consults with research partners about issues such as document digitization, ingest, description, and management, developing custom workflows, data models, deposit agreements, data management plans, and publication pipelines as necessary. Developed tools and techniques are available for other projects, thus enriching GAMS as well as the digital research environment for humanists.

- Personnel:
  Continuity of people is often correlated with the availability of infrastructure and data. In order to ensure long-term availability of the data as well as services, the centre invests in project staff for tasks such as software development, infrastructure management, processes, workflows, and content model design, as well as for document and metadata enrichment.

- Certifications:
  In a demonstration of our commitment for long-term preservation and to assure (potential) partners of this commitment, GAMS has undergone rigorous evaluation and has been certified as a trusted digital repository (since 2014), carries the CoreTrustSeal (since 2019), and is registered with the Registry of Research Data Repositories (ROAR).The team is currently working to certify GAMS repository as a CLARIN center.

4. Managing 100 DH legacy projects and building new ones: a pragmatic and holistic approach to archiving and sustainability (Arianna Ciula and Brian Maher, King's Digital Lab, King's College London)

King's Digital Lab's (KDL) (King's College London) contribution to archiving and sustainability practices in Digital Humanities (DH) will be presented along the following dimensions:

- (Human) Sustainability of expertise:
  As generational change occurs and in line with reorientations across the DH community (see Boyles et al., 2018), it has become increasingly clear that the surest way to sustainability is to ensure continuity of technical expertise, domain knowledge, and tacit understanding. KDL conceived and adopted a relatively flexible model with defined career development document and Research Software Engineering (RSE) role definitions (Smithies, 2019).

- (Technical) Sustainability of systems and technical stack:
  The second dimension needed to sustain the DH tradition and fulfil KDL's mandate to increase digital capability across the Arts & Humanities is caring for the cluster of technical systems comprised of hardware and software, web servers, network infrastructure, application frameworks, programming languages, tools (for project, data and code management), and equipment. In practice, sustainable management of lab projects required the adoption of limited server and development environment stacks, in a move away from the more flexible but difficult to manage environment used in earlier eras (for more details of the tools used to support the stack see https://stackshare.io/kings-digital-lab).

- (Operational) Post-project integrated in the lab Software Development Lifecycle:
  The techniques used to manage KDL rich and heterogeneous estate of legacy projects matured into an ongoing process of archiving and sustainability tailored to the Lab's historical, technical and business context. It is applied to new as well as legacy projects, in a manner that ensure systems as well as data are maintained

throughout defined life-cycles (King's Digital Lab, 2019). To control this, open ended Service Level Agreement (SLA) contracts are offered to Principal Investigators (PIs) of collaborative research projects to secure maintenance of legacy projects in their live state; however, other options for archiving are also possible and assessed (see also Smithies et al. 2019 and https://dev.clariah.nl/files/dh2019/boa/0648.html). To make the overall approach sustainable, it had to be integrated into the lab's Software Development Lifecycle (SDLC; see https://kingsdigitallab.github.io/sdlc-for-rse/), and in so doing align with KDL infrastructure architecture and core technical stack, while at the same time informing practices of forward planning for new projects.

KDL's contribution to the panel will reflect on how alignment across these three layers raises challenges but also poses the foundations for the sustainability of the Lab's ecosystem, hopefully offering a reference for others to reflect upon, adapt and improve.

5. **Keeping it Simple and Straightforward (Greg Newton, Stewart Arneil, Martin Holmes, Humanities Computing and Media Centre, University of Victoria)**

The University of Victoria long ago demonstrated its commitment to DH research by providing base-budget funding for the five-person Humanities Computing and Media Centre - a department in the Faculty of Humanities. As can be seen from the name, HCMC actually pre-dates the term Digital Humanities.

As a base-budget funded department, HCMC has the capacity to take on projects regardless of their level of funding - we regularly take on projects with no funding at all - and the commitment to support the project's outputs in perpetuity. This is only possible due to a critical mass of professors and executives seeing value over time.

For over twenty years HCMC has been consulting on and developing web applications in support of teaching and research. On behalf of our academic collaborators we work closely with library and systems colleagues who take primary responsibility for archiving and technical infrastructure, respectively. This is a strategic division of labour entailing ongoing communications with the benefits of specialization and scale.

Over the years we have come to recognize the inherent dangers of creating teetering stacks of complicated, fashionable technology that cannot stand the test of time. Experiments with several CMS's, JavaScript libraries, and so forth has invariably led us to the conclusion that the long-term cost of coping with breakage and security problems outweighs the short-term value these applications and libraries offer.

While we provide institutional support we are not keen on a never-ending cycle of upgrades and code-maintenance. To mitigate this we have become staunch supporters of KISS - in our case it might stand for "Keep It Simple and Straightforward". We take on very few projects that we did not develop, and when we do they are usually converted to a static site and archived or completely re-written.

Our Project Endings survey and interviews have made us doubly aware of the potential for catastrophe using technology that is not proven to be simple and durable. From our

perspective every project will benefit from adopting a KISS strategy, but perhaps especially those projects without institutional support.

References

Arneil, Stewart, Martin Holmes and Greg Newton. 2019. "Project Endings: Early Impressions From Our Recent Survey On Project Longevity In DH." Digital Humanities 2019 Conference, Utrecht, Netherlands. July 10 2019. https://dev.clariah.nl/files/dh2019/boa/0891.html

Boyles, Christina, Carrie Johnston, Jim McGrath, Paige Morgan, Miriam Posner, and Chelcie Rowell. 2018. 'Precarious Labor in the Digital Humanities – DH 2018'. In Digital Humanities 2018: Book of Abstracts / Libro de Resúmenes, edited by Élika Ortega, Glen

Worthey, Isabel Galina, and Ernesto Priani, 47–52. Mexico City: Red de Humanidades Digitales A.C. https://dh2018.adho.org/precarious-labor-in-the-digital-humanities/.

Breitenbücher, Uwe, Endres, C., Képes, K., Kopp, O., Leymann, F., Wagner, S., Wettinger, J., Zimmermann, M. 2016. 'The OpenTOSCA Ecosystem. Concept & Tools'. In: European Space project on Smart Systems, Big Data, Future Internet - Towards Serving the Grand Societal Challenges - Volume 1: EPS Rome 2016. SciTePress, pp. 112-130.

King's Digital Lab. 2019. 'Archiving and Sustainability | King's Digital Lab'. King's Digital Lab. 2019. https://www.kdl.kcl.ac.uk/our-work/archiving-sustainability/.

Neuefeind, Claes, Lukas Harzenetter, Philip Schildkamp, Uwe Breitenbücher, Brigitte Mathiak, Johanna Barzen, Frank Leymann. 2018. 'The SustainLife Project – Living Systems in Digital Humanities'. In: Proceedings of the 12th Advanced Summer School on Service-Oriented Computing (SummerSoC 2018) (IBM Research Report RC25681), pp. 101-112.

Neuefeind, C. and Schildkamp, P. and Mathiak, B. and Marčić, A. and Hentschel, F. and Harzenetter, L. and Breitenbücher, U. and Barzen, J. and Leymann, F. (2019). Sustaining the Musical Competitions Database. A TOSCA-based Approach to Application Preservation in the Digital Humanities. In: Book of Abstracts of the 29th Digital Humanities Conference (DH 2019), https://dev.clariah.nl/files/dh2019/boa/0574.html

OASIS. 2013. Topology and Orchestration Specification for Cloud Applications Version 1.0, http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.html

OASIS. 2019. TOSCA Simple Profile in YAML Version 1.2, http://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.2/TOSCA-Simple-Profile-YAML-v1.2.html.

Sahle, Patrick and Simone Kronenwett. 2013. 'Jenseits der Daten. Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner Data Center for the Humanities'. In: LIBREAS. Library Ideas #23, pp. 76-96.

Schildkamp, Philip and Brigitte Mathiak. 2019. 'Overview of life and death of digital scholarly editions' [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3483998

Smithies, James. 2019. ‘The Continuum Approach to Career Development: Research Software Careers in King’s Digital Lab’. King’s Digital Lab - Thoughts and Reflections from the Lab (blog). 7 February 2019. https://www.kdl.kcl.ac.uk/blog/rse-career-development/.

Smithies, James, Anna Maria Sichani, Carina Westling, Pam Mellen, and Arianna Ciula. 2019. ‘Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King’s Digital Lab’. Digital Humanities Quarterly. http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html.

Whitt, Richard. 2017. ‘Through A Glass, Darkly’ Technical, Policy, and Financial Actions to Avert the Coming Digital Dark Ages, 33 Santa Clara High Tech. L.J. 117. Available at: http://digitalcommons.law.scu.edu/chtlj/vol33/iss2/1