PRIDE PROPOSAL DESCRIPTION

PRIDE 2021 CALL

DEADLINE: 20 OCTOBER 2021; 14:00 CET

Title of PRIDE project	Deep Data Science of Digital History
Acronym	D4H
Name of DTU Coordinator(s)	Prof. Dr Andreas Fickers
Coordinating Institution	University of Luxembourg (UL), Luxembourg Centre for Contemporary and Digital History (C^2DH)

PRIDE PROPOSAL DESCRIPTION

1. DTU research programme

With the widespread digitisation of historical artefacts and the vast accumulation of born-digital content, scholars today have unprecedented computational access to our common cultural heritage, and this access will continue to grow. Despite the profusion of digital sources, most historians lack the methods and tools to interrogate these materials and fail to grasp the new opportunities (and limitations) of digital abundance. At the same time, experts in data science can now historically contextualise their research subject and develop new models that better explain both the past and the present. We strongly believe that these possibilities can be fully exploited via new modes of research that span the disparate fields of history and data science. More importantly, we propose to train a new cohort of PhD students who will become experts in both fields, capable of hybridising methods via a fully interdisciplinary approach, and to facilitate their integration into outstanding international networks by actively enabling research stays with renowned partners.

The key objective of D4H is to engage in a critical study of historical data by bringing together intellectual and technical resources generated across disciplines, particularly from digital history, social sciences and data science. To achieve this objective we propose to deepen the interdisciplinary collaboration between digital history and computer science by exploring the concepts of *deep history* and *deep data science*. Unlike traditional history, bound by access to mostly textual sources, deep history potentially interrogates the full span of human existence, extending backward beyond the written historical record and outward beyond ordinary archival sources. This groundbreaking approach requires new research infrastructures, including high performance computing and data management, as well as software and new algorithms for machine learning and information extraction, prediction of incomplete datasets and visualisation.

<u>Scientific Relevance and Epistemological Urgency</u>

We situate D4H among other newly established international training centres at postgraduate level that aim to facilitate collaboration between humanities research and data science to foster humanities-centred approaches to artificial intelligence and to provide infrastructures for this interdisciplinary exchange.[1] Although these initiatives facilitate cross-collaboration between data science and the humanities, we aim to accelerate and actively shape a dialogue that has already started between the fields, not least due to the successful work of DTU-DHH (see point 6). A commitment to interdisciplinarity is central for the

[1] See our partner institutions described below: ART-AI (Bath); Ghent Centre for Digital Humanities; Data Science Centre (Amsterdam); Digital Humanities Institute (Lausanne); Fondazione B. Kessler (Trento); History Lab (Columbia University).

development of a generation of AI-literate humanities scholars and hermeneutically trained data scientists, shaping and expanding the field of data science from a humanistic viewpoint (Drucker 2020).

Such mutual interest has already culminated in several exciting research projects at the European level with which we are associated, such as the Living with machines[2] project from the Alan Turing Institute in the UK, the European Time Machine initiative[3] and the SNF-funded *impresso* - Media monitoring of the past.[4] These flagship projects have paved the way for new forms of interdisciplinary collaboration between data curators, developers, data scientists, information specialists, web designers and historians. But when it comes to training and education, specialised programmes aiming at preparing the next generation of data-literate historians for the job are a rare exception. The recently established UKRI Centre for Doctoral Training in Accountable, Responsible and Transparent Artificial Intelligence at the University of Bath – a partner institution of our DTU – and a few PhD programmes in digital history in the United States (George Mason University, the University of Nebraska-Lincoln and the University of Virginia, with which the C²DH already has strategic partnership agreements) are currently the only academic institutions offering doctoral education in the field. D4H therefore aims to fill a critical gap in PhD training in a rapidly growing field and will as such offer interesting career opportunities for our students.

Conceptual Framework/ Research axes

We based our previous "Digital History and Hermeneutics" (DHH) DTU on two central concepts: the "trading zone" (Kemman 2021) and "digital hermeneutics" (Fickers 2020; Fickers and Tatarinov 2022). To reflect on the ongoing developments in the field of digital history, DHH was conceived as a space of experimentation where different epistemic cultures, disciplinary traditions and communities of practice could negotiate new forms of knowledge in the making (see point 6). For D4H, we plan to deepen our interdisciplinary collaboration and advance it to yet another level of understanding by focusing more explicitly on the trading zone between data science and history. In the process we expect to create new knowledge combining humanistic and scientific expertise in the interdisciplinary field of data science.

Key Research Questions

D4H addresses the following research questions that reflect the challenges of digital history based on a systematic analysis of "big data of the past" from a longue-durée perspective:

- How can various and different sources (from government statistics to photographic collections and geological and topographical investigations, to name but a few) be systematically linked, enriched, analysed, visualised and interpreted using a critical data management and curation framework?
- How can we make these large datasets and related analysis accessible to researchers across disciplines and to the general public interested in historical processes? Similarly, how can we perform these tasks in a way that triggers new research questions and innovative approaches based on traceability, shareability and reuse (FAIR principles)?
- How can we tackle the challenge of synthetic knowledge generation on a large scale through AI technologies and machine learning methods in an integrated way that guarantees that the integrity of the historic record is preserved and that the provenance of data can be retraced?
- How can modelling and simulation as technologies of knowledge production be practised following the intellectual tradition of critical hermeneutics? How can we identify patterns and structures in history and determine their (im)plausibility by comparing statistical evidence with historical relevance in a way that reflects the hermeneutic tradition of historical sciences?
- How can we ensure the sustainability, openness and transparency of research infrastructures, software and data, making them readable and understandable for future generations while also handling them ethically and preserving/documenting their context of creation, use and analysis?

Building on the conceptual and theoretical knowledge gained in the former DHH-DTU, we propose to organise D4H's activities along three complementary axes: 1) Deep Data & Knowledge; 2) Deep Analytics & Learning; 3) Deep Visualisation & Interpretation. These three axes will be complemented by a fourth transversal axis called Deep Time & History. This axis aims to link the epistemological and methodological debates faced by both historical and data sciences when dealing with "big data of the past" with more

[2] A cooperation with the British Library.

[3] Funded by EU Horizon 2020 and the Austrian Federal Government.

[4] Funded by the Swiss National Science Foundation (2017-2020).

fundamental questions of historical periodisation, multi-layered temporalities, and the interpretation of long-term historical processes and patterns.

Axis 1: Deep Data & Knowledge

Building upon insights gained for example in the new field of computational archival science (CAS),[5] the "Deep Data & Knowledge" axis aims to conduct an in-depth analysis of the **attributes, specificities, formats, histories and infrastructures of historical data**, and to better identify the challenges they raise for archivists, historians and data scientists. Indeed, digital "data" is never a given, never "raw", but always "cooked" (Drucker 2020; Gitelman 2013; Strasser and Edwards 2017). As such, digital datasets are the result of extensive and complex processes that transform gathered and curated information units into processable structures. Inevitably, this transformation comes with alterations and a potential loss of information – whether through the basic process of turning continuous signals into discrete ones, the choice of what to transform and what to leave out, or the creation of digital artefacts that may alter the original artefact. Central questions we aim to address in this axis are: How can we tackle the heterogeneity of data, their fluid or unstable nature in terms of volume, velocity, variety, validity, veracity and value (Lagoze 2014), in order to reconstruct past claims? How can we create a shared culture of data (and metadata) by mapping, analysing and understanding their shared components but also their specificities?

Five aspects are key to addressing these questions:

1) Using digitised and born-digital sources, big data or small data, rely on **a shared understanding of digital hermeneutics** but and imply being able **to put data in context** (e.g. their context of production, selection, heritagisation, recognising access bias, asymmetries, data frictions, etc.) (Borgman 2015).

2) Moreover, using data in historical research often requires the ability to work with several types of data that are not interoperable, may have various origins, and were created over several decades or centuries. It is therefore vital to document "data histories" and to develop skills in **data curation, management and governance** (Schiuma and Carlucci 2018).

3) **A deep understanding of data and datasets,** and the multi-layered transformations that data (and metadata as socio-technical constructions) may encounter during **their life cycle** (Kitchin 2021), is therefore a beneficial mutual outcome that computer scientists and historians should attain together. As (Crymble 2021: 176) suggests, curation within historical practice "involves selecting and contextualizing the contents of the collection". This deep insight into data may also help **to define and test criteria for data quality and integrity**.

4) It is important to **deconstruct digital knowledge representations** and **unbox data provenance,** and to **problematise the multi-layered architecture of datasets**. For example, historical maps are multi-layered carriers of knowledge combining text and images. Being able to identify (and potentially isolate) several layers of historical data in primary sources, compare them with other historical data and acquire a deep knowledge of these systems of the past is crucial (this is the focus of axis 3).

6) Finally, the Deep Data axis will explore new ways of digitising/preserving cultural artefacts using emerging technologies in order to extend the basis of available data by lowering economic barriers and **broaden the record by connecting dispersed datasets** with each other as part of the Luxembourg Time Machine initiative.[6]

Axis 2: Deep Analytics & Learning

In recent years, statistical modelling techniques, sensitivity analyses, predictive modelling and simulations have been successfully applied to pair historical research questions with the rigour of quantitative methods at scale. This in turn has encouraged deeper methodological reflections on their epistemological value and inherent challenges. Examples include the self-reflexive use of topic modelling to study the evolution of

[5] The new field of computational archival science (CAS) promises to address archival data curation challenges at scale (Stančić 2018; Marciano et al. 2019) and integrate computational methods into digital archival practice. These include creation, classification and retrieval. A crucial question here is to what extent domain expertise still plays a role in, for example, classification tasks, in which case human cultural/social etc. bias might be reproduced (Rolan et al. 2019) – and if not, how the algorithms were trained and what potential bias this might introduce (Trace 2021).

[6] The "Luxembourg Time Machine" is an initiative within the European Time Machine Organisation (see point 6). An interdisciplinary pilot project including 3 PhD students is currently financed by the Institute of Advanced Studies in connection with the "Audacity Programme" and co-funded by the Luxembourg Institute for Science and Technology (LIST).

notoriously elusive concepts such as literary genres (Underwood 2019), anti-modernism (Bunout and Van Lange 2019) or text reuse detection to reconstruct editorial practices based on newspaper content flows. Moving beyond a prior focus on textual sources, machine learning technologies can help detect objects in and similarities among images (Wevers and Smits 2020; Seguin 2018) and enable a data-driven analysis of the materiality of historical sources ("Numapresse" n.d.).

Machine learning methods and corresponding tools lend themselves well to expanding research beyond the scale of close reading by detecting otherwise unobservable patterns in large-scale data. Yet historical research practices have specific affordances since they are characterised by the twofold challenge of linking a **multitude of diverse information** available in different forms (text, image, sound, video) with **intangible expert knowledge**, while also compensating for often severe imbalances in the homogeneity, scale, quality and availability of the historical record (Poole 2017). What is more, historical research practices are characterised by their **exploratory nature**, which translates to the need to shift back and forth between research data gathering and data analytics. To accommodate this **iterative approach** to information gathering and the epistemological challenges inherent in analysing and interpreting historical data, novel approaches to **expert-guided data processing and method selection** need to be developed.

To address this and the aforementioned challenges and to achieve a more in-depth analysis of historical data and its affordances, we propose a "deep analytics" approach that focuses on six principles and methodological approaches for the integration of machine learning methods into historical research:

1) Generally, applications and analytics models need to be **cross-validated from a historical and data-science** point of view to determine the optimum fit between research objectives, available data and methodological choices. Specifically, researchers need to strike a balance between over- and underfitting historical data models. The trade-off between under- and overfitting remains one of the fundamental topics of machine learning research.

2) Researchers need to validate or reject machine learning outputs and thereby ingest their expert knowledge in the process, while also incorporating insights obtained during such interactions. **Human-in-the-loop** interactions facilitate the ingestion of intangible expert knowledge in deep learning processes to improve the performance of the latter, as exemplified by the human annotation of images (humansintheloop.org, (Budd, Robinson, and Kainz 2021). More recently, instances of **machine-in-the-loop** interactions have emerged. Here machine learning supports human reasoning, for example during the disambiguation of entity mentions (inception) or the semantic annotation of texts (DWISE). D4H aims for a dialectical relationship between human reasoning and machine outputs.

3) Missing data points are a recurrent problem in data-driven research (Ryan and Ahnert 2021; Blanke 2018). While missing data itself can not be generated synthetically, probable or plausible inferences can be made using **synthesised data** (Gelman and Hill 2007). To increase the accuracy of such artificially generated data, expert knowledge can be combined with machine learning. For example, a comparison of verified but known to be incomplete time series data with synthetically completed datasets may offer insights into the robustness of observed patterns and trends or indicate their volatility.

4) Comparative perspectives on data offer a multitude of opportunities to assess the significance of patterns observed in data through contextualisation (Düring et al. 2021). Comparing datasets of different sizes can be particularly challenging (Tantardini et al. 2019). Different datasets can also describe the same phenomenon albeit from different perspectives (Keyserlingk-Rehbein 2018). On the whole (visualisation-based, see axis 3), comparative perspectives offer a deeper understanding of the idiosyncrasies of the historical record, the data we extract from it and how it evolves over time, reveal (dis)similarities and help assess the robustness of observed patterns.

5) Knowledge discovery can be advanced with the **analytical and predictive power of networks**. The importance of networks in big data and data analytics is continuously increasing. Historical data can often be represented in the form of networks (Düring 2021). Dealing with data in the form of network structures reveals a huge amount of additional information that cannot be obtained exclusively with standard data analytics (Kerschbaumer et al. 2020; Düring et al. 2016). Furthermore, data transformed into network structures can provide evidence for missing information (Yang and Zhang 2016) and predict network dynamics (increasing predictive power) (Pan et al. 2016).

6) High-performance computing (HPC) facilities process large-scale datasets or run simulations which require capacities beyond those of ordinary computer hardware. HPCs have so far been used somewhat rarely for humanities data. Examples include image similarity detection for historical paintings (Seguin 2018) or newspaper images (Wevers and Smits 2020) or the processing of large-scale audio collections (NESS-SoundSynthesis). Their extremely powerful computational capacities can also be used to analyse

or traverse large-scale complex networks with millions or even billions of nodes and exponentially more links between them.

Axis 3: Deep Visualisation & Interpretation

Emerging technologies in the fields of information visualisation and human-computer interaction have not only altered our interpretation and understanding of the past; digital tools have also empowered historians with new ways of thinking and communicating ideas (Liu 2018). Working with digital tools and digitised visual materials offers clear benefits to any scholar or student of history. However, there is often a missing link between how these tools are currently used and how they could empower historians (Ayers 2018). But without a critical understanding of the logics of data visualization, historical interpretation will remain speculative. Deep interpretation means not only being able to explain or show our understanding to others, but also reframing and revisiting our own beliefs and hypotheses, and translating historical arguments into "graphic arguments" (Drucker 2020) and new forms of temporal and spatial sampling of historical information (Manovich 2020). But as previously mentioned, most of the machine learning mechanisms that power these state-of-the-art techniques are black boxes, i.e. the values within them are subsymbolic and it is often not possible to understand why a given decision was taken (Broussard 2018). This problem becomes particularly challenging when the decisions made by these mechanisms are wrong or intriguing (Broussard 2018). Against this backdrop, the recent advent of eXplainable AI (XAI) has made it possible to probe and peek into black-machine learning mechanisms such as deep neural networks. XAI provides explanations for various aims including: understanding the underlying mechanisms of ML, debugging, controlling and updating, and training.

This is a particularly challenging task, for five main reasons:

1) **Explanations provided by XAI need to be personalised and context-aware**, since the explanation required by one type of user (e.g. a field researcher) might be different from what is expected by another (e.g. an archivist). A single user might also need different explanations in different contexts. This means that the XAI mechanism needs to construct a model of the context and to tailor the explanations accordingly, thereby enabling researchers consuming the explanation to assess their understanding, probe the system and update their knowledge if necessary.

2) **Providing these explanations means bridging the gap between symbolic and subsymbolic AI.** This is a persistent gap that has split the AI field into two branches for more than half a century. The first approach, known as "connectionism", tries to emulate neurons in human brains, the second is the traditional symbolic AI, which is mainly interested in knowledge modelling using logic and reasoning. The emergence of XAI has helped bridge the long-standing dichotomy in AI since recent XAI research aims to provide explanations for heterogeneous AI mechanisms, allowing symbolic reasoning to be conducted on top of subsymbolic approaches.

3) One of the most challenging issues when developing XAI mechanisms, especially in cross-disciplinary contexts, is **measuring how good an explanation is,** respectively **how valid an interpretation is.** A good explanation based on XAI mechanisms does not make a valid interpretation in history. As Stephen Ramsay argued in his book *Reading Machines* in 2011, we should try to "locate a hermeneutics at the boundary between mechanism and theory (...) that we channel the heightened objectivity made possible by the machine into the cultivation of those heightened subjectivities necessary for critical work" (Ramsay 2011: X)

4) After the underlying decision model (i.e. the machine learning mechanism) is explained, **the system should help the user detect any biases and enforce fairness.** This issue has been outlined in the literature, since many black-box ML mechanisms can give biased decisions. The origin of this bias can be traced mainly to skewed training data or bias in the decision-making itself. Explainability has been proposed as an effective method to help reduce this bias by exposing it and allowing the user to opt for a fair decision.

5) Finally, while data visualisation facilitates the interpretation of data gathered by humans, we argue that the **semantics of the data have to be made explicit for both humans and machines**. Ontologies and knowledge graphs are semantic resources that have been used to integrate datasets (Brouwer and Nijboer 2017) and, via a combination of reasoning and ML techniques, have allowed the discovery of new knowledge from these datasets (Hoppe, Dessi, and Sack 2021). However, as illustrated in axis 1, the maintenance and evolution of these ontologies and knowledge graphs are crucial for accurate interpretation and have been identified as particularly challenging (Sentance 2017).

Transversal Axis: Deep Time & History

"Time" as a word and as a concept is used in all scientific disciplines, but with enormous variability in scale, precision and weight (Hunt 2008). As "time" is a central concern of historical research, history can offer a unique contextualisation of this concept for other disciplines (Hall 1980). Despite the particularity of history, and more generally despite differences between disciplines when they measure, study and interpret time, the concept is the central shared denominator on which this DTU is based. More specifically, by combining computer science, digital humanities and history, D4H aims to explore the concept of time by investigating, broadening and redefining the notion of ***deep time.*** As a concept used since the 1980s to refer to *geological time* (McPhee 1998), deep time has since been appropriated, discussed and broadened in order to reflect on continua, patterns, intangible phenomena and shorter cycles (Shryock and Smail 2011). Deep time will therefore be understood as a concept enlightening geological time (Lyle 2015), anthropocene time (Chakrabarty 2018; Irvine 2020), but also the Braudelian *longue durée* (Braudel 1958; Lamouroux 2015; Armitage and Guldi 2015) and all time elements and temporal strata that are embedded into data and computational thinking. This axis is transversal to the three others, as (deep) time serves as a central criterion for analysing and understanding historical contexts, data, structures (and infrastructures) and developments.

Time can be linked to the traditional chronological ways of representing history – a ceaseless succession of seconds, minutes, hours, days, months, years… but this linear concept of time has never satisfied historians, for whom history is not just "one damn thing after another", an apocryphal quote which neatly summarises what history is *not*. Time in history can instead be understood as perceptions of time by a great variety of historical actors; it is therefore not linear, but a complex assembly of perceptions and projections. One of the most famous essays on time and history is Fernand Braudel's work on *longue durée* (Braudel 1958). Braudel developed a multilayered notion of time and temporalities: short term (political history), mid-term (economic history), and long term (*civilisation matérielle*). Braudel's thinking, deeply rooted in the intellectual tradition of the *Annales* school (Burke 2015), has been highly influential in post-war historiography; his aim was to build bridges between different disciplines in an attempt to understand structural patterns, economic cycles and mental dispositions as well as socio-cultural traditions of people and societies. On a more theoretical level, reflections on the different temporal strata, layers and semantics of historical time have been systematically developed in the works of Reinhart Koselleck (Koselleck 1989; 2003; 2011) and more recently enriched by the concept of "regimes of historicity" by (Hartog 2003).

Thinking about time, temporal regimes, and multilayered temporalities as well as space in a broader and longer framework of historical analysis that better captures the potential of digital research has been at the heart of the *History Manifesto* by Jo Guldi and Davird Armitage (Guldi and Armitage 2015). *The History Manifesto* provoked intense resistance and debate ("Special Issue: 'La Longue Durée En Débat'" 2015). So far, cliometrics and cliodynamics remain largely ignored in the field of digital history (Hudson and Ishizu 2017). Because research practices continue to move inexorably in the direction of digitization, we believe that an enormous untapped potential for a fruitful encounter between quantitative and qualitative approaches in historical research remains (Lemercier and Zalc 2019). In confronting the rich historical literature on social, economic, or mental structures with data-driven methods of pattern-detection using AI and machine learning techniques, the DTU aims at addressing crucial epistemological questions about the relationship between statistical evidence and historical relevance, singularity and causality, and historical change and continuities (Van den Akker 2018; Törnberg and Törnberg 2018).

This transversal axis therefore aims at turning the metaphor of "deep time" into a productive heuristic instrument by problematizing the complex notions of multi-layered temporalities both in a "horizontal" (longue-durée) and "vertical" (superimposed temporal regimes) perspective. The focus will lie on three key aspects:

(1) **The temporalities that are inherent to "data".** Most knowledge graphs operate in a *radical now* with no conception of time and therefore often fall short in providing the contextual information they are meant to offer. How have/are data shaped by past and current states of the art of computer science and archiving technologies? How can digitised material and born-digital heritage be seen as encapsulating multiple temporalities and practices (embedding several layers of time during their creation, when they are made available, used, reused, etc.)?

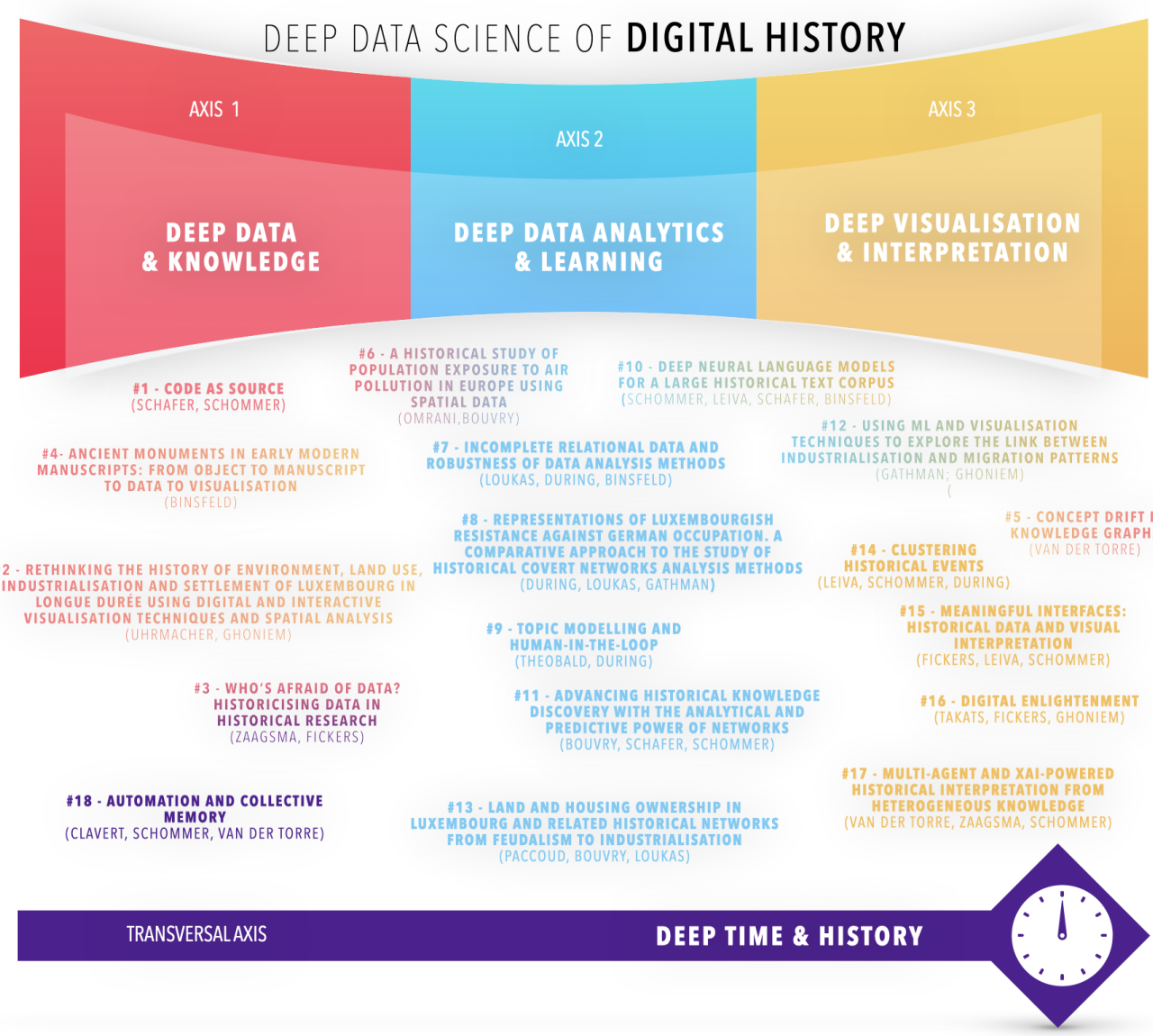
(2) **The extent to which data fit historians' definitions of time**: How are data and data uses in history evolving in relation to both the research questions of their time and also available tools, methodologies, etc.? How can we transform traditional primary sources into data that help us

understand multilayered time and understand and model the simultaneity of the non-simultaneous? The question of forgotten or hidden data that are rediscovered or promoted in a new way through time is also key.

(3) **Patterns and cycles, continuities and disruptions** and the way data may (or may not) mirror these changes through time: how can we model temporalities? How can we move from retrospective modelling to dynamic forms of simulation, enabling historians and data scientists to "play" with historical parameters and potentially to extrapolate past patterns into the present and future? Can computer-generated models and simulations help to identify and visualize "critical moments" or "transitions" in different temporal regimes (environmental, economic, political) and help to grasp multiple causalities and conflicting temporal logics?

List of indicative topics for PhD Dissertations

The following diagram shows a list of potential PhD topics spread over the different axes. For a more detailed description of the role of the post-docs and their topics see the section "Interdisciplinarity and inter-institutionality". Flexibility will be given to supervisors and PhD students in determining the research agenda and identifying suitable co-promoters:



Legal and Ethical requirements

The handling of issues related to data protection will be addressed on a scientific level in axis 1, on a practical level by informing all candidates of ethics and data protection issues for individual research, and also on an institutional level (see BnL partner workshop). Participants involved in D4H must comply with Luxembourgish and European regulations on the protection of personal data, especially with EU regulation

679/2016, known as the General Data Protection Regulation (GDPR), with regard to any processing of personal data that may be carried out in connection with the project. The University adopted and communicated to all staff the Data Protection Policy that must be respected. The team will consult the Data Protection Officer about the projects involving personal data processing and ensure to:

• protect the integrity and confidentiality of any personal data processed in connection with the project.
• respect the confidentiality of personal data of employees involved in the performance of tasks.
• anonymise/pseudonymise any collected data if required or requested by participants, and obtain approval from the UL Ethics Review Panel for any data collection.
• fill in the records of personal data processing and check whether a Data Protection Impact Assessment is required.

Contribution of supervisors to current state of the art of relevant research in the field of DTU

The D4H team is the result of a careful combination of different disciplinary traditions and expertise in a number of relevant topics. The **team of historians** involved in this new DTU represents a great variety of expertise in terms of thematic specialisations, methods and experience in digital history and draws on strategic recruitments in the field of digital history in recent years. The C^2DH boasts a critical mass of digital history experts that it would be hard to find at any other academic institution in the world. The team consists of experts in digital hermeneutics (Andreas Fickers: 2020, Gerben Zaagsma: 2013), software and tool development for historians (Sean Takats: Zotero, Tropy), web archives and born-digital heritage (Valérie Schafer: Musiani et al. 2019), social network studies (Frédéric Clavert: 2021), historical network analysis (Marten Düring), text mining and topic modelling tools (Frédéric Clavert, Marten Düring) and text encoding and human-computer interaction (Florentina Armaselu). This know-how is complemented by the expertise of two colleagues from the Institute for History specialising in digital cartography and mapping (Martin Uhrmacher: Uhrmacher, Kass, and Pauli 2021) and 3D technologies in digital archaeology (Andrea Binsfeld: 2010).

The second pillar of the DTU is built around a **group of computer science and data sciences**, mostly based at the Interdisciplinary Laboratory for Intelligent and Adaptive Systems (ILIAS) in the Department of Computer Science at the University of Luxembourg. The five professors from ILIAS are Pascal Bouvry, special advisor to the Rector for high-performance computing and Head of the University's High-Performance Computing Centre (Paseri, Varrette, and Bouvry 2021); Leon van der Torre, an expert in the field of knowledge representation, logic and multi-agent systems (Dong, Markovich, and van der Torre 2020); Christoph Schommer specialised in data science and machine learning applications (Guo, Höhn, and Schommer 2019); Martin Theobald, an expert in big data technologies and analytics (Fletcher and Theobald 2019); Luis Leiva (Leiva and Vidal 2012; 2013), a specialist in human-computer interaction.

A central ambition of the DTU is to reach out to other leading research centres in Luxembourg outside the University. By joining forces with the "Data Science and Analytics" research group at the Luxembourg Institute for Science and Technology (**LIST**), the consortium considerably strengthens its expertise in the fields of human modelling, knowledge representation and symbolic artificial intelligence, as well as in interactive visualisation and geocomputation. Dr Mohammad Ghoniem (a specialist in the visualisation of multilayer dynamic networks (McGee and et al., 2019) and Dr Cédric Pruski (an expert in the evolution of ontologies and knowledge graphs (Cardoso, Da Silveira, and Pruski 2020) will bring extensive expertise in data visualisation. In addition, D4H will benefit from digital expertise in the field of social sciences by teaming up with colleagues from the Luxembourg Institute for Socio-Economic Research (**LISER**). One PhD candidate will be supervised by Prof. Christina Gathmann (Head of the Labour Market Department and affiliated Professor of Economics at UL, a labour economist with a focus on migration, (Gathmann and Keller 2018), and the history of political and economic institutions as well as a strong interest in data science. In addition, Dr Antoine Paccoud, a social geographer with a keen interest in the inequalities created through land and housing ownership, will supervise work on the history of Luxembourgish real estate by analysing digitised property transactions (Paccoud 2020). Finally, Dr Hichem Omrani a specialist in machine learning, modelling and simulation of complex spatial systems (Zięba-Kulawik et al. 2021), will contribute expertise on the effect of air pollution on society and public health by comparing serial environmental data from the past with present-day pollution measurements.

For more detailed descriptions of areas of expertise, please see attached narrative CVs.

Partner institutions and added value of collaboration

Contracting Partners

The DTU-D4H encompasses the three major research institutions in Luxembourg conducting doctoral training: the **University of Luxembourg (UL)** with **LIST** and **LISER** as contracting partners. In addition to renowned experts in the field of digital history and computer science affiliated with the UL, LIST will contribute with extensive expertise in applied data science. The team is experienced in conducting multi-institutional projects in the field of digital cultural heritage (see BLIZAAR, PI: M. Ghoniem). They also have close links with industrial players and provide training expertise in the field of business management and intellectual property. LISER contributes quantitative environmental, economic and administrative data and expertise on how to analyse and interpret the data. Its focus on Luxembourgish history from a sociological and geographical point of view links D4H research to pressing societal issues in Luxembourg such as migration, pollution, health and housing. Its extensive experience with doctoral training networks such as the DTU ACROSS (PI: F. Docquier) and the Marie-Curie ITN on the urban health system (led by M. Dijst) is of great value for the DTU-D4H.

Institutional Partners

The urgent need for cross collaboration between data science and digital history in terms of research and training is reflected in the immensely positive feedback from institutional partners on a global scale that are keen to collaborate with D4H in this field.

Our **training partners** will link D4H PhD candidates to world-class experts in the field and commit to providing attractive opportunities for international exchange.

The UKRI Centre for Doctoral Training in Accountable, Responsible and Transparent Artificial Intelligence at the University of Bath aims at developing world-leading AI via an explicitly interdisciplinary approach. The Data Science Centre at the University of Amsterdam facilitates data-driven research. Similar to D4H, its focus is on providing training in AI for other fields via an interdisciplinary doctoral-level approach. Our American partner History Lab, Columbia University (USA) is aggregating large corpora of US and international documents using machine learning and natural language processing techniques. It offers exchange, advanced training and a valuable link to the American digital humanities community. With a focus in the areas of geo-spatial humanities, semantic web technologies and deep mapping, the Ghent Centre for Digital Humanities is renowned in the Benelux region. It provides training in data standards, tools and linked data. The Digital Humanities research group at the Fondazione Bruno Kessler in Trento, Italy, is leading in the field of textual data processing and the use of AI in historical investigation in multilingual digital archives. DARIAH European Research Infrastructure Consortium, Paris will facilitate workshops with a focus on network-building and professionalisation for future leading researchers in the field of digital humanities. The German Historical Institute, Paris will support D4H with its expertise in science communication and open access. It runs the blogging infrastructure hypotheses.org. The Digital Humanities Lab at the EPFL has considerable expertise in the customisation and application of computational tools and software for big data analysis and visualisation and initiated the Time Machine flagship project.

Our **Luxembourg partners** will support us in linking research output to the general public and the industrial sector in Luxembourg. They provide infrastructure, resources, technology and knowledge of current national developments and digital policies in the field of digitalisation and cultural heritage. LuxProvide S.A. is a company providing high performance data analytics and AI solutions on an international scale as Luxembourg's high-performance computing (HPC) centre. Joint workshops will address the basics of HPC and how to use the supercomputer for research. The Bibliotheque nationale de Luxembourg (BnL) is providing us with access to Luxembourg's heritage collections. The BnL is involved in a large-scale ongoing digitisation project and is continuously adapting its practices. Workshops are planned on copyright issues, data modelling and application development. The Archives nationales de Luxembourg is the

biggest holder of governmental and state records in Luxembourg, systematically conserving and inventorying. It has extensive expertise in corpus identification and metadata description technologies. The Centre National de l'Audiovisuel (CNA) is committed to sharing its expertise in data management and digitisation of audiovisual cultural heritage collections and is an established partner for public outreach activities.

Our **research partners** will provide the D4H team with specialised knowledge and resources for the entire DTU or individual projects. The Time Machine Organisation, Vienna, links more than 600 research institutions worldwide to build a large-scale simulator mapping 5,000 years of European history, transforming an unprecedented volume of archives and large collections from museums into a digital information system (see point 1, axis 1). Professor Dr Johanna Drucker, University of California, who has conducted groundbreaking research on visualisation, digital hermeneutics and interpretation, will support PhD and postdoc projects in research axis 3 and the transversal axis in visualising and interpreting multiple chronologies and timescales. Software Heritage (INRIA) is the largest publicly available open archive of software source code, with more than 11 billion unique source files collected from a variety of platforms, from over 100 million projects and will support research in axis 1. The School of Natural & Built Environment (SNBE), Queen's University Belfast, will be our main scientific partner for the geospatial aspects in deep mapping and cultural heritage visualisation, historical GIS surveying, digital analysis and landscape characterisation. The German Archaeological Institute (DAI) in Berlin is involved in building a large-scale digital collection of research materials from excavations and developing a modular database system for the documentation of field research projects, supporting projects with open and FAIR research data.

Interdisciplinarity and inter-institutionality (distribution of PhDs and post-docs)

D4H is a large-scale collaboration with the ambition of creating a research and training hub for interdisciplinary research in data science and history in Luxembourg. The principal investigator will be Prof. Dr Andreas Fickers, Director of the C^{2}DH. The PI will be supported in coordinating this broad collaborative network by the post-doc researcher Dr Juliane Tatarinov, who can draw on extensive skills in academic project management and science communication. She will conduct participant observation research on the DTU as an interdisciplinary "trading zone", focusing on the emergence of interactional expertise, shared language and knowledge brokerage (Fickers and Heijden 2020). Her post-doc position will be funded on an equal basis by the FSTM (2 years) and the C^{2}DH (2 years), illustrating the importance assigned to successful project management and academic reflection on this unit by the main partners. Her work will be split into 50% project management and 50% research activities. The post-doc researcher funded by LIST (supervised by Dr Mohammad Ghoniem) will investigate how multiple experts work together to analyse a corpus of textual and non-textual data (images and maps) and explore and evaluate alternative interaction and visualisation methods based on large high-resolution displays, e.g. interactive tabletops and wall-sized displays, to cater for historical research needs. The main outcome of the project will be an interactive visualisation dashboard. The post-doc researcher funded by LISER and coordinated by Prof. Dr Aline Muller will specialise in historical geography and/or economics. S/he will serve as a link between computer scientists and big data infrastructure (HPC) and investigate the specifics in spatio-temporalities of data (planned for the transversal axis "deep time & history").

Starting date will be January 2023. DTU-D4H involves 18 PhD positions and 4 post-docs (post-docs funded by LIST, FSTM, C^{2}DH and LISER) in the fields of history, data science, social geography and economics.

- Faculty of Science, Technology and Medicine (FSTM): 6 PhDs and one post-doc focusing on epistemological questions in the first half of the programme
- Luxembourg Centre for Contemporary and Digital History (C^{2}DH): 6 PhDs and one post-doc focusing on epistemological questions in the second half of the programme
- Faculty of Humanities, Social Sciences, and Education (FHSE): 2 PhDs
- Luxembourg Institute for Science and Technology (LIST): 1 PhD and 1 post-doc to create synergies on data visualisation

- Luxembourg Institute of Socio-Economic Research (LISER): 3 PhDs and one post-doc bridging social sciences and data sciences

After the onboarding of PhD students, we will set out the principles of collaboration, the value of working in an interdisciplinary environment, research goals and envisioned output in a project charter in the very early stages of the programme, as is best practice (Ahnert 2019).

2. Training and career development

The D4H PhD candidates will be trained as AI-literate, humanities-educated specialists in either computer science or digital history, with dedicated training in the other, complementary discipline. The training will be divided into (1) training provided by the doctoral schools, mainly addressing transferable skills, (2) D4H in-house interdisciplinary training and career development and (3) specialised training by external partner institutions.

(1) Doctoral schools and doctoral programme

The PhD students in the DTU-D4H will be enrolled into the doctoral schools at the University of Luxembourg, including the PhD candidates from our contracting partners LIST and LISER. Ten PhD students (C^2DH, FHSE, LISER) will be affiliated to the Doctoral School in Humanities and Social Sciences (DSHSS). The doctoral training offered within the DSHSS includes specialised modules in disciplinary training, interdisciplinary training, and transferable skills training. The D4H training programme will be included in the online course programme and ECTS will be assigned to respective DTU courses. Eight PhD students (FSTM, LIST, LISER) will be enrolled in the Doctoral School in Science and Engineering (DSSE). The DSSE offers 7 programmes, including the Doctoral Programme in Computer Science and Computer Engineering (DP-CSCE) with the main research areas Communicative Systems, Intelligent & Adaptive Systems, Security & Cryptology and Software & Engineering. Activities are focused on transferring scholarly skills, identifying original approaches and solutions, conducting interdisciplinary research, and teaching and communicating with target groups, reflecting the skill set required for personal career development. The system of training modules is structured similar to the DSHSS, and D4H training courses will be implemented here as well.

(2) D4H specific training

Central to our training approach is the **weekly lab space** to encourage **trading zone discussions** on terminology, methods and tools. This will be complemented by lectures from external experts throughout the duration of the programme. The handling of ethically sensitive issues as well as the establishment of a data management & protection plan will be a core concern during the launch phase (see point 1).

Skills training and multiliteracies: The training programme aims to create a common language for the disciplines involved (see point 6). The expertise of the consortium will be used to provide an introduction into central questions, methods, and approaches in digital humanities as well as an introduction into historical thinking and practices aimed at computer scientists. Conversely, training on data gathering, analytics and reasoning will be provided for the humanities scholars. As part of the quality assurance system for doctoral education, a research and training plan (RTP) must be drawn up by the doctoral researcher and supervisor(s). Next to general courses which aim to familiarize the group with the epistemic traditions of the different disciplines, we will identify taylored training needs of the two cohorts in a second training phase. Here, skills training will be demand-driven in order to provide a tuned skill set for each of the PhD students.

Career development: Students can take advantage of the career development environments at UL, LIST and LISER, where there is a growing budget for training early career researchers. Transferable skills offered include career development, lecturing and teaching, proposal writing, presentation and communication skills, project management, academic best practices, ethics in research, open access and open data, as well as entrepreneurship and design thinking. LISER provides a set of empirical research training such as data and survey methodologies and policy expertise. LIST offers expertise in intellectual property, creating spin-offs and business skills in cooperation with Luxinnovation. Our main focus will be

on fostering interdisciplinary and international mobility with the aim of building an extensive, high quality professional network. Our partner institutions in the private, public and academic sectors offer multiple training and exchange opportunities. For students targeting an academic career we work closely with leading doctoral training centres in Europe and the US to provide students with multiple opportunities for global networking and embedding their research (e.g. DARIAH ERIC's scholarly events). We also offer various teaching opportunities in related fields to help doctoral candidates boost their CVs. For PhD students targeting a career in industry, the LIST team (in cooperation with Luxinnovation) will provide training in this area covering project management, communication skills, intellectual property rights management, and entrepreneurship. LISER offers a specific career development plan including workshops and individual coaching sessions for career advice with experienced external career coaches.

(3) Specialised training from external partner institutions

One key component of the programme is the international network into which we are embedding our research and training. The training partners will open up specialised courses on tools, methods and best practices for our candidates and offer links to renowned experts in their institutions. We are proud to offer internships and research stays at the following partner institutes: University of Bath (ART-AI group), the DH Lab at Columbia University, GHI Paris, ICT group Fondazione Bruno Kessler in Trento, Ghent Centre for Digital Humanities, the School of Natural & Built Environment (SNBE) at Queen's University Belfast, and the Data Science Centre at the University of Amsterdam. The table below describes the core training programme offered by D4H in addition to what is offered by the doctoral schools and institutional training. It comprises regular research seminars, a lecture series, specific training courses on basic concepts and skills, a series of lectures given by international renowned scholars in the field, external research and training offerings, two off-site retreats for PhD students and supervisors as well as a masterclass for PhD candidates for individual feedback and a final conference to create synergies.

Training Activity	Aim and content	Lead organiser
Trading Zone Discussions and Lectures		
research seminar, lecture series (monthly format, years 1-3)	Regular space for informal discussions on shared questions & problems, preliminary results Lectures by renowned external scholars	All DTU-D4H candidates and invited external guests
Off-site retreats (years 2/3)	Mutual learning on innovative topics and new approaches (2-3 days) and team-building	All DTU-D4H candidates and supervisors
Skills training/multiliteracies (tailored to the RTP)		
Workshops, lectures (year 1: basic concepts and overview; year 2: advanced)	**Digital history:** Text mining, sentiment analysis, data curation and management, interface criticism, source criticism and digital hermeneutics, data quality assessment, historical network analysis, archival & information science; Advanced R: web scraping, text analysis and image recognition	C^2DH (During, Zaagsma, Schafer, Clavert, Fickers, Wieneke, Takats), LISER (Gathman)
	Data science: Basics in machine learning and artificial intelligence, machine learning in historical networks, data aggregation, sensitivity analysis, regression analysis, predictive modelling, simulation analysis, graph neural networks and deep learning, human-computer interaction, data science for humanities, big data analytics, knowledge representation and reasoning	DCS (Theobald, Leiva, van der Torre, Schommer), LISER (Omrani), LIST (Ghoniem, Pruski)

	Mapping Time and Space: Meaning of time from a historical and geographical perspective, historical GIS and history of cartography; deep mapping, meaning of time as a complex system, econometrics; quantitative analysis and statistics, introduction to Python and R; data conceptualisation, cleaning, manipulation and visualisation; GIS and R, understanding coordinate systems, creating maps, geo referencing historical maps, extracting pixel information	LISER (Gathman, Omrani, Paccoud), FHSE (Binsfeld, Uhrmacher), C^{2}DH (Clavert), DCS (Bouvry), with partners: Lilley (Belfast)
Career Development		
Workshops (selection), years 3/4	Path 1 industry sector: Project management, communication and presentation skills, intellectual property rights management, entrepreneurship	LIST, Luxinnovation
	Path 2: academic career, public sector: Academic teaching, proposal writing, sustainable data, open science, research ethics, policy evaluation, research integrity, digital sustainability	C^{2}DH (Takats, Schafer, Sillaume), Binsfeld, DARIAH workshops; LISER (Zana-Nau)
Networking event, year 4	Job fair with active players from the public and private sectors, PhD students will have the opportunity to engage with future employers	With guests such as Luxinnovation, Digital Luxembourg, the EU Commission, Luxembourg ministries
Specialized training by and with external partners		
Partner workshops (selection), years 1 and 2	Digital collections	Carlo Blum (BnL)
	Visualisation and hermeneutics	Johanna Drucker (UCLA)
	Use of the HPC cluster	Luxprovide
	Science communication; open science	Mareike König (GHI, Paris)
	Deep learning and image recognition	Melissa Dell (Harvard)
Research and training abroad, year 2 or 3 Internships with partner institutes (< 3 months), external summer schools	Aim: to foster international mobility, further accumulate specialised knowledge, create new incentives for training and research, create visibility and possibilities to embed the candidate's research in an international framework and build up an academic network	University of Bath, Columbia University, GHI Paris, Fondazione Bruno Kessler in Trento, Ghent Centre for Digital Humanities, Queen's University Belfast, University of Amsterdam

Masterclass, year 3 (2 days)	To create synergies across the projects and to collect individual feedback	All DTU-D4H candidates with international partners
Final conference year 4 (3 days)	To discuss the case studies, synergise initial results within an international discourse and offer training in conference organisation	D4H focus groups

3. **Ability of the host institution(s)/DTU to manage PhD training and quality of supervision**

Profile and qualification of the coordinator

The coordinator of the DTU, Prof. Andreas Fickers, has held the Chair in Contemporary and Digital History at the University of Luxembourg since September 2013. As Director of the Luxembourg Centre for Contemporary and Digital History (since 2016), he is member of the Management Team of Luxembourg University and participated in the "Digital Strategy Group", one of three university think tanks initiated by the Rector to develop the University of Luxembourg into a model research university for the 21st century. He is actively engaged in developing strategies to promote digital literacy and skills in teaching and research (for both students and academic staff) on a centre, faculty and University-wide level. Fickers has an outstanding track record of international publications and long-standing experience in the management of international research networks and projects (see CV). He has initiated and maintains key management positions in several European research networks and serves as an advisor on multiple academic and editorial boards. Part of his work has led to the creation of open access, internationally peer-reviewed journals (www.viewjournal.eu; https://journalofdigitalhistory.org/en). Fickers has proven his qualities as a project manager in a number of projects at national and European level in which he developed necessary skills for the successful management of interdisciplinary projects, including organisational talent and responsibility, budgetary management, intellectual inspiration and rigour, and social and leadership skills. All of these projects were driven by the ambition to combine academic excellence, intellectual creativity and methodological innovation with a clear output-oriented publication and dissemination strategy. As head of DARIAH-LU, he is promoting the development of a digital research infrastructure for the humanities in Luxembourg. He also serves as a member of the CLARIAH International Advisory Panel in the Netherlands. Alongside his role as head of the DTU-DHH, he codirects the "International Interdisciplinary History" Trinational Doctoral School together with Prof. Hélène Miard-Delacroix (Sorbonne University) and Prof. Dietmar Hüser (Saarland University) and has recently received the "Outstanding Mentorship" Award from the Luxembourg National Research Fund (FNR).

Organisation and management of the DTU

With regard to the coordination of the DTU we can draw on extensive experience from the predecessor DTU-DHH. Useful tools have been the implementation of a **management board** with regular meetings (every two months) to discuss the strategic basis for training and research activities, organisational issues and any corrective measures required. The board will represent all institutions involved and all career stages (supervisors, post-docs and PhD candidates). For day-to-day communication and coordination of the multi-institutional consortium and interdisciplinary PhD cohort, **a coordinating post-doc as a contact point** for D4H-related questions and initiatives with regard to research, training, organisation and exchange will be essential. The main partners, the C^2DH and the FSTM, will therefore provide funding for such a position for the duration of the programme. The C^2DH will furthermore provide a software platform to support **project management** in a transparent and consistent way throughout the programme. The coordinating post-doc will support and monitor the successful implementation of the training programme and maintain relationships with the partners, while the post-docs at LISER and LIST will be contact points for inter-institutional communication. We will gladly take up the FNR's suggestion to appoint **DTU champions** to address important aspects of our collaboration, such as improving on diversity and gender equality, addressing sustainability in procedures and purchases and monitoring the well-being of the team.

Esther Zana-Nau (LISER) and Hervé Muller (C^2DH) are trained mediators who will be tasked with resolving conflicts and attending to mental health issues within our team.

Interaction in the group will be centred around an informal **weekly DTU brown bag lunch** and an **slack channel** to foster formal and informal everyday communication. We follow a strict open science policy and will maintain GitHub pages to share code and tutorials and create a **DTU blog**, as for DTU-DHH. **Working groups** will be set up for specific tasks such as organising a lecture series or editing the website. This proved to be effective and socially engaging in the previous DTU-DHH. **Social activities** such as excursions, get-togethers during conferences and off-site retreats will also help to strengthen team spirit and foster identification with the interdisciplinary approach.

Recruitment

Our recruitment strategy will focus on overcoming the gender and diversity gap that is still very much present in all disciplines. We are aware of the social responsibility to address gender equality in our recruitment strategy as a top priority, especially given the strong male bias at supervisor level. We will collaborate with UL Gender Equality Officer Skerdilajda Zanaj, who will provide coaching during the hiring process, and D4H will participate in the Mentorship Programme organised by the Gender Equality Office. Following the communication of PRIDE decisions by the FNR, the consortium will liaise with UL, LIST and LISER HR Departments to launch the recruitment process for doctoral candidates; this will be an open, transparent and merit-based process as set out in the European Code of Conduct. Applicants will be considered in one cohort (vacancies will be published in summer 2022 via EURAXESS and reputed portals). The Steering Committee will draw up a shortlist from the list of applicants. Admission is subject to legal criteria, the availability and interest of a thematically relevant supervisor, the candidate's academic ability and research potential, and the suitability and feasibility (including financial) of the PhD research project. Shortlisted candidates will be invited for online interviews. The Steering Committee will collectively select the best candidates and will either make an offer or invite them for on-site interviews (autumn 2022). The successful PhD candidates will be invited to UL induction days. Upon admission, they will be informed about their rights and duties, and those of their supervisor(s).

Quality control, supervision and evaluation, assessment

All candidates benefit from the quality assurance of the UL. A central Office of Doctoral Studies provides and applies general procedures across the doctoral schools', and provides support and guidance to the doctoral candidates. The quality assurance in UL's doctoral education is achieved by means of several instruments:

Supervisors and ADR: All supervisors must have a PhD and the necessary supervision skills and rights (ADR – *Autorisation à diriger des recherches*), as stipulated in the University Act and Internal Regulations. ADRs can be granted to both internal and external eligible PhD holders through an evaluation procedure involving 3 internal and 3 external experts. New ADR holders receive mandatory supervision training. All D4H supervisors will be encouraged to complete this training. The **Thesis supervision committee** (*Comité d'encadrement de thèse* – CET) is composed of three members holding a doctorate and is encouraged to include foreign partners. The role of the CET is to guide the candidate during the research programme towards the objectives of the PhD in annual meetings, discussing achievements and difficulties encountered and potentially proposing corrective measures, and determining the next steps. Cross-disciplinary PhD projects and cross-faculty collaboration are key to this DTU, and each CET should ideally reflect this interdisciplinarity by creating co-tutorships based on the principle of cross-field supervision. The **doctoral education agreement (DEA)**, which sets out and regulates the rights and obligations of the doctoral candidate, the supervisors and the CET members and lays down the specific conditions for a given doctorate. The **research and training plan** (**RTP**), produced by the doctoral candidate and the supervisor. It provides details about research and training components including the main research questions and how they will be addressed. The RTP also describes how the mandatory 20 ECTS will be acquired and any additional requirements. In the yearly presentation to the CET, candidates report on the status of their research and the future research plan as well as any completed and planned training. The **thesis defence** defence is the traditional conclusion of the quality assessment of a doctorate. The defence jury, composed of both external and internal experts, evaluates the thesis based on the criteria laid down in Article 37(6) of the University Act: a) academic knowledge; b) research autonomy and relevance of the research methods used; c) structure of the work and literature; d) quality of the written presentation of the thesis; e) quality of the defence of the thesis. All manuscripts are evaluated twice for plagiarism and

published on Orbi.lu. **Feedback after completion of doctoral studies** is organised to identify any issues regarding supervision or systemic deficits that require corrective action.

Care for the individual

In the event of disagreements concerning doctoral studies or any situations of conflict, a mitigation process is provided in the DEA, with several levels of intervention: (1) DTU mediators, (2) DTU management team, (3) Doctoral School, (4) Vice-Rector for Research, who may obtain a final settlement. An Ombudsperson is accessible to all doctoral candidates. For psychological support, doctoral candidates can contact the university psychologist and the service for well-being and inclusion. The UL has implemented a policy on research ethics. If conflicts and questions about research integrity arise, the Luxembourg Agency for Research Integrity (LARI) can be contacted to perform an independent enquiry and investigation. The Litigation Committee deals with conflicts regarding formal and legal matters.

Research environment and infrastructure

The DTU-D4H will be embedded into institutions that have already been involved in coordinating a DTU and have established quality processes for doctoral training (see above). As an integral part of lively and emerging research centres and UL departments, the DTU-D4H offers a high quality, personalised research and training environment that encourages intellectual interaction and supports social activities.

4. Contribution to the strategic goals of the involved institution(s)

The DTU-D4H is fully in line with the strategic goals of the C^{2}DH, UL-FSTM, UL-FHSE, LISER and LIST. For this reason, 2 post-doc positions will be funded by UL, 1 by LISER and 1 by LIST to invest and create synergies in digital history and data science and to harness the full potential of digitalisation in accordance with national research priorities.

C^{2}DH: The C^{2}DH has developed into the leading research institution in the field of digital history in Europe. Its long-term strategic ambition is to co-design and operate a global open-science platform for history in the digital age and to shape the critical debate on contemporary history by promoting a hands-on digital hermeneutics approach. Based on ongoing flagship projects such as the Digital History and Hermeneutics DTU, the FNR/PEARL project DHARPA (Digital History Advanced Research Projects Accelerator) and the FNR/ATTRACT project PHACS (Public History as the New Citizen Science of the Past), D4H will boost the C^{2}DH's position as a key player in data-driven historiography. By training the next generation of data-literate history students and exploring new forms and formats of data-driven publishing (see *Journal of Digital History*), the C^{2}DH aims to promote a new culture of academic recognition for innovative data-driven scholarship in the field of history.

FSTM: The University adopts a collaborative approach that aims to strengthen and promote multidisciplinarity through various mechanisms, such as the FNR funding scheme PRIDE. Digital transformation is also one of the key areas identified for the future of the University. The Department of Computer Science (DCS) is committed to meeting this challenge and collaborating with others to achieve research excellence. DCS research covers topics that span both fundamental and practical aspects of computer science, and currently focuses on four broad research areas: Communicative Systems (ComSys), Intelligent and Adaptive Systems (ILIAS), Software Systems (LASSY) and Cryptography Security (LACS). Ongoing collaboration with digital humanities is a thriving interdisciplinary research strand within the department and with ILIAS in particular.

FHSE: The main disciplines covered by the Department of Humanities at the University of Luxembourg (FHSE) are history, language and literature studies, arts and media studies, multilingualism, parliamentary studies and philosophy/ethics. Research in all these fields is closely linked with Luxembourg society. In methodological terms, digital humanities approaches, combined with a critical assessment of digitalisation in general, are becoming increasingly crucial for research and teaching. The Faculties Centre of Ethics and Digitalisation will contribute to the imminent ethical issues resulting from the ongoing boom in digitalisation and artificial intelligence.

LIST: D4H is in line with LIST's strategy and the Data Science and Analytics Unit. DSA carries out impact-driven research by combining computers with human capabilities, drawing on topics like knowledge representation and reasoning as well as visualisation to make better decisions. D4H will benefit from ITIS Department investments like the Data Analytics platform, which comprises an HPC infrastructure, a "cognitive analytics pillar" implementing data analytics, artificial intelligence and big data solutions, and an

interactive visualisation wall based on multiple synchronised data views. LIST has invested in the creation of a dedicated unit within the HR Department that focuses on doctoral training and preparing researchers for future careers.

LISER: To further address social challenges and continue to elucidate and solve complex interdisciplinary and cross-sectoral societal problems, LISER has adopted two methodological investment plans for the 2022-2025 period: a) the integration of large-scale databases in combination with advanced data science methods and computing capabilities; b) experimentation and close collaboration between science and public authorities to accelerate the implementation of appropriate public policies. LISER will continue to invest in its expertise in microsimulation, GIS, machine learning and behavioural and experimental economics, and is aligned to the HPC cluster. The DTU DRIVEN, an interdisciplinary data-driven computational modelling group and the University's Master of Data Science programme are driving forces in strengthening collaborations and partnerships.

5. Outcomes of the DTU

Scientific and educational impact: D4H aims to publish 18 PhD theses, more than 60 peer-reviewed papers, one special issue in an open access peer reviewed journal, and about 60 presentations at international conferences. These targets take different dissemination and publication strategies in the different fields into account. PhD graduation requires 3 publishable papers in computer and social sciences, one monograph and at least 2 peer-reviewed papers in digital history. D4H aims to run 45 skills training and career development seminars for PhD candidates, 10 partner workshops, 20 guest lectures, 2 off-site retreats, 1 masterclass, 1 final conference, 1 networking event and 1 outreach fair.

Dissemination, open access and open data: D4H supports open access publishing, including - where possible - open data access. Dissemination of results in open access scholarly journals and at conferences is required and transfer to a lay public is encouraged. UL has developed an outreach and IPR guide, and employs a specialist to support researchers in this area. There will be specific training for PRIDE doctoral candidates on open access and open data strategies for researchers. All tangible D4H outputs will be published on UL's OrbiLU open access server. Temporary or partial embargoes are possible. Given the strong emphasis on data use for research, LISER and LIST will train PhD students in developing data management plans based on effective, GDPR-compliant techniques for documentation, data storage and archiving.

Economic, technological and societal impacts: D4H engages in a critical study of historical data based on a reflexive approach to technology and digital infrastructures – a highly relevant issue for a data-driven society and economy. We are convinced that the interdisciplinary and cross-institutional training of PhD students in D4H will have a lasting impact on policy development in the areas of cultural heritage, data sustainability, data cycles, access to knowledge, and the societal and economic impact of AI technology in Luxembourg and beyond. Finally, the DTU will eventually become the intellectual and scientific backbone of the Luxembourg Time Machine initiative and produce a new generation of data-literate humanities, respectively humanistically trained data scientists that are high on demand on the academic and professional job market.

Outreach activities: D4H will work actively to disseminate research results to the general public, as well as to institutions and decision-makers in the field of digital education, digital heritage and AI/digital literacy for the humanities in the international academic environment. The D4H team will organise one Forum Z on "Big data of the past" (facilitated by C^2DH), addressing citizens and national stakeholders. The D4H blog will host reflexive essays and encourage the publication of podcasts and video-essays by the students. The variety of D4H topics also offers potential for established FNR science communication formats such as "chercheurs à l'école", "researchers days", or LUX:plorations comics. LISER has received FNR RESCOM funding for its research seminars and hosts the outreach series cafés-débat on "science & société", where D4H topics will be included. A "Luxembourg Time Machine" hackathon will be organized in collaboration with the European "Time Machine Organisation" (TMO).

Career prospects for the PhD candidates are excellent and depend on their research focus and professional skill set. Digital historians can pursue a career as an academic scholar in digital history and digital humanities, with opportunities offered by the extensive D4H partner network promoting post-doc job opportunities and other academic avenues. They can also be employed as social science and humanities engineers or data journalists, or in careers in heritage institutions dealing with born-digital heritage (e.g. in our partner institutes BnL, CNA and AnL). For PhD candidates specialising in quantitative analysis (supervised by LISER and LIST), jobs in socio-economic research institutes, national or European statistics offices and regional agencies for entrepreneurial training or urban and development planning are

attractive career paths. Computer scientists will have ample opportunities to be employed in academic projects at international level or as data scientists in industry, public agencies or political institutions. Depending on the adaptability of applications developed during the research and on IP law, the creation of spin-offs is another potential career path.

6. Comments on follow-up proposal for "Digital History and Hermeneutics" (DTU-DHH)

The Digital History and Hermeneutics Doctoral Training Unit (DTU-DHH) was based on two central concepts: the concept of "trading zone" and the concept of "digital hermeneutics". The trading zone concept, emanating from history of science and sociology of knowledge, inspired the design of the DHH as a collaborative space of knowledge production in which methodological interdisciplinarity and theoretical bricolage formed the mental framework for critical debate and discussion (Collins, Evans, and Gorman 2007). DHH consisted of historians, philosophers, computer scientists, geographers, information scientists and experts in human-computer interaction, who collaborated in an interdisciplinary setting characterised by experimentation, creative uncertainty and appropriation of new tools and methodologies for digital history research. This diversity of disciplines and approaches was mirrored by the diverse range of "sources", "documents" and "data" under study, ranging from textual data to oral testimonies, toponymies, pictures, material objects, archaeological data, and computer models.

Inspired by the call of Fred Gibbs and Trevor Owens "to publicly experiment with ways of writing about their methodologies, procedures, and experiences with historical data as a kind of text" (Gibbs and Owens 2011), we encouraged our PhD students to reflect on the "usage" of historical data not simply as evidence or as "self-identical" (Drucker 2012) but from multiple viewpoints and based on the principles of the hermeneutic theory of critical interpretation. In this sense, DHH approached digital history as what Julie Thompson Klein refers to as "deep interdisciplinarity": a mode of collaboration that can alter disciplinary practices and create new hybrid languages (Klein 2015: 142). The large and diverse scholarly output of DHH is both impressive and telling: 13 PhD theses, 15 peer-reviewed papers and articles, an edited volume, 35 conference presentations and 2 apps, one of which was runner-up for Application Impact award at the 28th International Joint Conference on Artificial Intelligence (S. Haddadan: https://www.ijcai19.org/demos.html). DTU-DHH researchers regularly published reflexive blog entries and were involved in several (co)teaching activities.

Succeeding DHH, D4H both builds on lessons learned (Fickers and Heijden 2020; Kemman 2021) and adds new training elements and collaborative activities to enhance interactional expertise and successful knowledge brokerage. Since the creation of a "forced trading zone" during the "DH incubation phase" of DHH showed limited success in producing a "shared vocabulary" - in fact disciplines were too wide-ranging - D4H will focus on building a trading zone between "just" two fields/disciplines: history and data science. While DHH looked at the challenges of digital history from a humanities perspective, D4H aims to adopt **a more balanced perspective**, exploring the methodological and theoretical challenges of data science when confronted with often complex, diverse and non-structured humanities data. Based on the assumption that both fields actually struggle with very similar challenges owing to the implications of digitisation and datafication on their methodological toolkits and conceptual apparatuses, D4H will tackle issues of data management, data criticism, data visualisation and data interpretation within the joint framework of digital hermeneutics. Inevitably, this requires a serious intellectual and communicative investment by all partners involved, including supervisors, external experts and doctoral students.

Another new element for D4H is a **joint project** to which all DTU students and supervisors can contribute despite their different fields of expertise and disciplinary backgrounds: the **Luxembourg Time Machine**. D4H will build on two ongoing initiatives related to the Time Machine project: the pioneering LuxTIME project, funded by the UL Institute for Advanced Studies, and the FNR-funded INITIATE project "Luxembourg Time Machine", which aims to build a national consortium to analyse and interpret historical "big data".