Time Machine: Big Data of the Past for the Future of Europe

Deliverable D2.2
WP2 (Pillar 1) Roadmap

Preamble:

This deliverable is part of a structured set of outputs produced and built upon during the 12-month TM CSA project. As such, the content contained within was further refined, synthesised and improved throughout the project and in particular when combined with material from other deliverables during the production of the full D8.5 TM LSRI Strategy and Implementation Proposal. Please be advised that the most up-to-date version of any information found in this document will be found in D8.5, where it can also be viewed in proper context as part of the entire TM LSRI proposal.

Abstract

A road map is presented for the Science and Technology actions that constitute Pillar 1 of the Time Machine LSRI. The objective is to develop cutting-edge computational methods, specially through AI, to access, organise, and understand large-scale cultural heritage collections. The targeted technology achievements will allow virtual time traveling by extracting knowledge and establishing links over space and time.

For this purpose, a taxonomy of technologies relevant for the Pillar is defined, divided in three main areas “Data”, “Computing and AI” and “Humanities and Social Sciences”. Each of these areas is further divided in subareas and individual topics. This modular structure makes it possible to design interrelated actions for multidisciplinary work groups across Europe to radically transform large-scale humanities studies, data archives, user interfaces and the way the past is analysed, in order to improve our understanding of our future.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 820323

Project Identification

Project Full Title	Time Machine: Big Data of the Past for the Future of Europe
Project Acronym	TM
Grant Agreement	820323
Starting Date	1 March 2019
Duration	12 months

Document Identification

Deliverable Number	D2.2
Deliverable Title	Science and Technology (Pillar 1) Roadmap
Work Package	WP2
Delivery Due Date	31 October 2019 (Month 8)
Actual Submission Date	19 December 2019 (Version 1.1) 7 February 2020 (Version 1.2) 21 February 2020 (Version 1.3) 09 March 2020 (Version 1.4) 05 June 2020 (Version 1.5)
Leading Partner	FAU, UA
Main Author	Andreas Maier, Gustavo Fernández Riva, Mike Kestemont, Emma-Louise Silva
Author Contributions	Alicia Fornés, Andrea Torsello, Andreas Kuczera, Andy Stauder, Antoine Isaac, Bagdan Ludwiczak, Bjoern Eskofier, Bruno Vallet, Bénédicte Bucher, Carmen Brando, Christoph Lofi, Christophe Sidonie, Claartje Rasterhoff, Deborah Papiernik, Djordje Kuzmanovic, Dorit Raines, Ellen van Noort, Erich Schubert, Erik Buelinckx, Fabian Hollaus, Fabian Wolf, Fabio Vitali, Fauzia Albertin, Finn Dammann, Florian Niebling, Francisco Casacuberta, Frank Bauer, Georg Glasze, Georg Thallinger, Gernot A. Fink, Giorgos Tolias, Guillaume Touya, Hanan Salam, Harald Koestler, Hendrik Herold, Ilja Van Damme, Ingo Zechner, Iza Romanowska, Jaqueline Klusik-Eckert, Javier de la Rosa, Jeroen Deploige, Joan Andreu Sánchez Peiró, Josep Lladós, Jouni Tuominen, Julien Perret, Katerina Hornickova, Klaus Meyer-Wegener, Konstantinos Kotis, Krzysztof Kurowski, Lars Wieneke, Marieke Van Erp, Marcus Breitenstein, Markus Wirth, Martin Boß, Mathieu Brédif, Michelle Lindlar, Milosz Ciznicki, Nanne van Noord, Nathalie Abadie, Pawel Wolniewicz, Peter Bell, Peter Uhrig, Robert Hecht, Robert Sablatnig, Sander Münster, Stefan Evert, Thierry Poibeau, Thomas Bähr, Tim Soens, Tomasz Parkola, Valérie Gouet-Brunet, Vassilis Papadakis, Vincent Jolivet, Werner Bailer
Non-Author Contributions	Tomasz Slava, Harald Sack, Johan Oomen, Maria-Esther Vidal, Didier Richard, Michael Loebenstein, Erwin Verbruggen, Fulvio Rinaudo, Iason Jongepier, Jan Van Gemert, Martin Winter, Marc Pierrot-Deseilligny, Arianna Traviglia, Bianca Arellano, Anastasios Kesidis, Alexander Mehler, Eleni Galiotou, Gerhard Weber, Jouke

	Verlinden, Wolfgang Nejdl, Kuldeep Singh, Laurent Costa, Blanca Arellano, Jan van Gemert, Marc Stamminger, Enrico Gobbetti, Wolfgang E. Nagel, Gerben Zaagsma, Andreas Fickers, Stéphane Lamassé, Frederic Clavert, Sean Takats, Jaap Evert-Abrahamse, Reinout Rutte, Kaarel Sikk, Susanne Muth, Raphaelle Lapotre, Torsten Schlauwitz, Daniel Röwenstrunck
Reviewer(s)	Frédéric Kaplan, Julia Noordegraaf, François Ballaud

History of Changes

Date	Version	Author	Comments
31.10.2019	1.0	Andreas Maier, Mike Kestemont, Gustavo Fernández Riva	Document ready for submission, after integration of comments by internal reviewers.
17.12.2019	1.1	Andreas Maier, Mike Kestemont, Emma Louise Silva	Reviewer comments addressed, Annex C added.
07.02.2020	1.2	Andreas Maier, Mike Kestemont, Emma Louise Silva	References added and minor changes implemented.
26.02.2020	1.3	Andreas Maier, Mike Kestemont, François Ballaud, Emma-Louise Silva	Alignment with RFC Tree
05.03.2020	1.4	François Ballaud, Frédéric Kaplan	Minor changes on milestones/RFC implemented
03.06.2020	1.5	Kevin Baumer	Added preamble to title page which points to D8.5, per final review meeting.

Disclaimer

This document expresses only the authors' views. The European Commission is not liable for any use that may be made of the information contained therein. Moreover, this document does not express the opinion of European Commission and does not in any case replace the European Commission documentation.

Table of Contents

INTRODUCTION

This document describes the roadmap for Pillar 1 “Science and Technology” and details the broader vision underlying it.[1] In order to organise and facilitate the scientific and technological challenges associated with the Big Data of the Past, Pillar 1 adopts a modular, layered structure, consisting of a series of interdependent modules, each advancing and producing concrete results at their own pace. For this, we rely on a scientific **taxonomy** (see below) that differentiates between various relevant **subdomains** in science and technology, while not forgetting the interdisciplinary cross-dependencies between them. There are three main areas in this taxonomy: (1) **Data**, (2) **Computing and Artificial Intelligence** and (3) **Social Sciences and Humanities**, which broadly correspond to the areas designed in the Pillar’s original Scoping Document (“Data”, “Computing” and “Theory”)**.** Each of these areas has a specific aim and contribution in relation to the TM:

(1) **Data**: Enable persistent digital access to more than 2000 years of linked historical data;

(2) **Computing and AI**: Develop generic methods to explore, connect, and simulate historical information;

(3) **Social Sciences and Humanities**: Provide explanatory models of historical evidence that lead to new, plausible narratives, radically transforming the manner in which SSH engages with and interfaces with the past.

Each of these three areas is divided into a series of sub-fields with their own disciplinary traditions, methodologies and long-and mid-term goals. In order to implement and shape the developments outlined in these documents, the model of publication known as **Request for Comments** (RFC) will be essential. RFCs are freely accessible publications to establish rules, recommendations, and architectural choices. In most cases, the discussions and results of work carried out in Pillar 1 will take up the form of RFCs. The details of that implementation are discussed in the Roadmap for Pillar 2 “TM Operation”.

The **overall objective** of this Pillar can be summed up as follows:

Overall objective: To develop cutting-edge computational methods, enhanced with Artificial Intelligence, to access, organise, and understand large-scale cultural heritage collections. This technology will enable virtual time travelling by extracting knowledge and establishing links over space and time. We aim to bring multidisciplinary work groups in Europe together to radically transform large-scale humanities studies, (archival) data processing, user interfaces and the way we analyse the past to understand our future.

The **logic** behind the structure of this document is as follows: we start by critically surveying the state of the art in each domain. Next, we go on to identify the domain-specific challenges in each domain which the TM will address (targeted achievements). The proposed methodology is

[1] Appendix A provides an overview of the overall approach for the design of the Time Machine Large Scale Research Initiative, of which Pillar 1 is a constituent part.

detailed in the next section of the actual roadmap, in which a series of realistic milestones are proposed to be solved via a call for proposals cycle. Finally, key performance indicators and stakeholders are identified and discussed in light of this section.

The present document engages in a close dialogue with the roadmaps for other Pillars, in particular that of Pillar 2: the targeted achievements in Pillar 1 directly feed the exploitation plans of Pillar 2. More specifically, this involves the creation of the Time Machine Infrastructure and the Official Components (see the *Roadmap for Time Machine Operation,* by Pillar 2). Throughout this document, we shall clarify in various places in which Pillars 2 and 3 must interface and where common protocols and performance indicators must be shared. The most relevant, recurrent issues in this respect shall include:

1. **Technical charter**: *Which project-wide formats, protocols and infrastructure for data creation, storage, and exchange must be negotiated by the consortium members?*
2. **Digitisation Hubs**: *Which technology must be developed to realize hotspots of local digitisation initiatives, the results of which can be seamlessly aggregated into a pan-European CH data infrastructure?*
3. **Digital Content Processor**: *Which standard processing techniques and generic routines will become available in the next 10 years for the automatic processing of digital CH content?*
4. **Time Machine Data Graph**: *Which scientific and technological challenges must be addressed to realize a distributed super-computing and storage system, which can be efficiently indexed, searched, updated and exploited?*

RESEARCH AND INNOVATION PLANS

State of the Art

To realize the Pillar's overall objective, TM must interact with a variety of domains in science and technology, which each come with their own methodological traditions and discipline-specific challenges. In this section, we provide a critical assessment of the present-day state of the art in these domains. Below, these fields are discussed and reviewed individually, relying on a clear-cut taxonomy which was developed in preparation of this roadmap, in order to identify the areas in science and technology which can be expected to be most relevant for the TM. While the project generally emphasizes the importance of interdisciplinary work, this subsection follows a fairly conventional taxonomy for the sake of clarity. Additionally, the taxonomy follows the overall three-branch structure of the Pillar's subdomains, i.e. Data, Computing and SSH.

Taxonomy of Relevant Areas in Science and Technology (Pillar 1)

1. DATA

- **1.1. Data Acquisition**
  - 1.1.1. 2D digitisation
  - 1.1.2. 3D digitisation
  - 1.1.3. Audio digitisation
  - 1.1.4. Film and video digitisation
  - 1.1.5. Scientific analysis
- **1.2. Data Modelling**
  - 1.2.1. Knowledge Modelling
  - 1.2.2. Data formats
  - 1.2.3. Metadata Formats and Mapping between Standards
  - 1.2.4. Annotation
- **1.3. Long Term Preservation**
  - 1.3.1. Bitstream layer
  - 1.3.2. Functional layer
  - 1.3.3. Semantic layer
  - 1.3.4. Trustworthy archives

2. COMPUTING AND ARTIFICIAL INTELLIGENCE

- **2.1. Computer Vision and Pattern Recognition**
  - 2.1.1. Text recognition
  - 2.1.2. Graphic document processing
  - 2.1.3. Image processing and analysis
  - 2.1.4. Indexing and Retrieval
  - 2.1.5. Understanding and Interpretation
  - 2.1.6. Recognition and Detection
  - 2.1.7. Person, Face Identification
  - 2.1.8. Modelling, Registration, and Reconstruction
  - 2.1.9. Audio recognition & transcription
- **2.2. Natural Language Processing**
  - 2.2.1. Methods for Resource Scarce Languages
  - 2.2.2. Orthographic normalisation and variation handling
  - 2.2.3. Machine reading / Document understanding / Question answering

2.2.4. (Structured) Metadata extraction, manipulation, and translation/mapping
2.2.5. Discourse analysis

2.3. Machine Learning and Artificial

2.3.1. General Artificial Intelligence
2.3.2. Supervised Learning
2.3.3. Unsupervised Learning
2.3.4. Weakly Supervised Learning
2.3.5. Transfer Learning
2.3.6. Deep Learning
2.3.7. Universal Representation Space
2.3.8. Explainability
2.3.9. Bias / Fairness / Ethics in AI

2.4. Human-Computer Interaction and Visualisation

2.4.1. User-centred Interfaces
2.4.2. Access to large-scale information retrieval and recommender systems
2.4.3. Virtual / Augmented / Mixed Reality
2.4.4. Accessibility and Learning, Adaptive, and Cognitive Interfaces
2.4.5. Motivational Design
2.4.6. Big data visualisation
2.4.7. User Experience
2.4.8. Virtual research environments

2.5. Computer Graphics

2.5.1. Rendering
2.5.2. Animation
2.5.3. Immersive, Virtual, and Augmented Reality
2.5.4. Interactive Computer Graphics and Computer Games
2.5.5. Procedural Content Generation

2.6. Super Computing

2.6.1. Scaling and distribution
2.6.2. Dynamic provision of computing platform
2.6.3. Cloud computing
2.6.4. Secure distributed computing

3. SOCIAL SCIENCES AND HUMANITIES

3.1. Theory

3.1.1. Qualitative vs. quantitative studies: resistance and acceptance
3.1.2. Increase research scope in SSH
3.1.3. Simulation studies
3.1.4. Digital methods

3.2. Disciplines

3.3.1. History
3.3.2. Language and literature
3.3.3. Archaeology
3.3.4. Art history & media studies
3.3.5. Geography and demography
3.3.6. Musicology
3.3.7. Digital humanities
3.3.8 Urban studies

Data

1.1. DATA ACQUISITION deals with the technologies necessary to digitize and model CH Objects, extending to cities and territories within the scope of Time Machine. Applications of current approaches range from objects such as paintings, transparencies or various written documents, which are conventionally digitised via 2D technologies, to more voluminous objects, such as statues, buildings and landscapes, for which 3D digitisation approaches are more common.

As for **2D Digitisation (1.1.1),** several programs exist, produced at different times and places, with a heterogeneous coverage of objects.[2,3] Digitisation benches equipped with digital camera backs are the main technique for this large-scale task and, although they vary greatly, most tend to be manually operated or at least manually assisted. Traditional flat-bed scanning is used on a smaller scale. The technology for high quality digitisation has become more cost effective through the years. The biggest advance in the last couple of years is the use of smart live image processing tools: almost no further post-processing is needed. For opaque material state-of-the-art digitisation standards include the Dutch Metamorfoze guidelines, IMPACT Centre of Competence recommendations for digitisation projects and the FADGI of Library of Congress (USA).

3D Digitisation (1.1.2) is a newer field, with the third dimension to various meanings involving different types of information including, but not limited to, the temporal, spectral or structural domain.[4] Moreover, the typology of 3D digitizable objects are really varied in terms of scaling: landscapes, archaeological sites, architectures, sculptures, paintings, books, etc. Multiple methods have been developed, although none of them has yet succeeded to become a routine procedure, especially for historically important objects.[5] A list of the currently available 3D acquisition techniques includes structured light, stereo vision, digital holography, photogrammetry,[6] many different kinds of spectral imaging,[7] as well as X-ray tomography,[8] ultrasound, Lidar,[9] etc.

2D and 3D digitisation are essential for the TM, as most of the relevant objects in the Big Data of the Past fall under these categories. In particular, fast and large-scale acquisition is an important goal in TM, a goal that can be achieved using distributed and cheap solutions such as the Scan Tent or highly automated scan robots. Therefore, improvements of image to text (e.g. OCR, HTR) or image tagging techniques will be of high importance during the coming years. However, when considering the more recent history of **Audio (1.1.3.), Film, and Video (1.1.4)**

---

[2] Digitization program of the Biblioteca Vaticana, Città del Vaticano. http://www.digitavaticana.org/
[3] World Digital Library program, Library of Congress, Washington. https://www.loc.gov/collections/
[4] Smithsonian 3D Digitization program. https://dpo.si.edu/programs
[5] Adamopoulos E., and Rinaudo F., 3D Interpretation and fusion of multidisciplinary data for heritage science: A review. In: *27th International CIPA Symposium*, 42, pp. 17-24. (2019).
[6] McCarthy J., Multi-image photogrammetry as a practical tool for cultural heritage survey and community engagement. In: *Journal of Archaeological Science*, 43, pp. 175-185. (2014).
[7] Fischer C., and Kakoulli J., Multispectral and hyperspectral imaging technologies in conservation: current research and potential applications. In: *Studies in Conservation*, 51: sup1, pp. 3-16. (2006).
[8] Stromer D., et al., Browsing through sealed historical manuscripts by using 3-D computed tomography with low-brilliance X-ray sources. In: *Scientific Reports*, 8.1, pp. 1-10. (2018).
[9] Schindling J., and Gibbes C., LiDAR as a tool for archaeological research: a case study. In: *Archaeol Anthropol Sci*, 6, pp. 411-423. (2014).

digitisation is also relevant: each domain has a set of very specific problems and techniques. Video and audio carriers are deteriorating rapidly and the original playback equipment needed to carry out the digitisation is becoming obsolete.[10] Film digitisation is very labour-intensive and it's very likely that new technologies will be developed in the upcoming years.[11] Automated metadata acquisition, like speech to text, is therefore also expected to be improved in the near future.

Aside from traditional CH objects, such as documents, paintings, sculptures, and buildings, the TM will also digitize bigger geographical features like cities, landscapes, and territories, as well as other environmental information in relation to the climate. Since the ancient times, man has proposed methods to measure its environment and the Earth, and has invented geodetic coordinate systems as well as maps. With the digital era, the earth, our societies, and climate have been digitised more or less systematically by cadastral and mapping agencies, geological surveys, meteorological institutes, and statistical surveys into different products, generating digital elevation models, topographic databases, land use products, and virtual reality models. These kinds of objects pose particular challenges for digitisation, challenges that current technology is addressing inadequately.

Finally, data acquisition is not merely limited to creating digital images, but also to extracting some of their invisible physical and chemical features,[12,13,14,15,16] a process known as **(1.1.5) Scientific Analysis**, which is extremely important in ensuring correct preservation and restoration of objects, as well as sustainable management of our environment and the places where we live.

1.2. DATA MODELLING describes the conceptual framework, the technological approaches and the knowledge representation tools necessary to support and harmonize the digital forms of cultural artefacts so as to allow for fruitful interoperation, which translates into handling primary, secondary, and tertiary data as well as annotations. Primary and Secondary Data (objects and their metadata) should be accessible via persistent identifiers and preserved by well-established institutions (e.g. Cultural Heritage Institutions) in a trustworthy Digital Archive/Repository (the TM Box). Additionally, research infrastructures should make it possible to interconnect such data by means of Linked Open Data technologies and semantic enrichment. All data should be stored in FAIR repositories which enable data submission and retrieval in a standardised way, enabling transparent data processing and transmission pipelines, and ensuring streamlined data migration between repositories.

---

[10] Bigourdan J. L., and Reilly J. M., Effectiveness of storage conditions in controlling the vinegar syndrome: preservation strategies for acetate base motion-picture film collections. In: *Joint technical symposium* (2000).
[11] de Jong A., *Digital Preservation Sound and Vision: Policy, Standards and Procedures*. Netherlands Institute for Sound and Vision (2019).
[12] Artioli G., and Angelini I., *Scientific methods and cultural heritage: an introduction to the application of materials science to archaeometry and conservation science*, Oxford University Press (2010).
[13] Cucci C., Delaney J. K. and Picollo M., Reflectance Hyperspectral Imaging for Investigation of Works of Art: Old Master Paintings and Illuminated Manuscripts. In: *Accounts of Chemical Research*, 49.10, pp. 2070-2079. (2016).
[14] Ruberto C., et al. Imaging study of Raffaello's "La Muta" by a portable XRF spectrometer. In: *Microchemical Journal*, 126, pp. 63-69. (2016).
[15] Petrucci F., Caforio L., Fedi M., et al., Radiocarbon dating of twentieth century works of art. In: *Appl. Phys. A* 122**,** pp. 983. (2016).
[16] Re A., Albertin F., Avataneo C., et al., X-ray tomography of large wooden artworks: the case study of "Doppio corpo" by Pietro Piffetti. In: *Heritage Science*, 2, pp. 19. (2014).

The field of **Knowledge Modeling (1.2.1)** deals with uncertainty and trust, provenance, layered annotations, multidimensional annotations, graph models, and multilingual data. These are complex tasks, as secondary and tertiary data about the same artefact are intrinsically multiple in number and often inconsistent or contrasting in content. This is, however, not meant as a limitation, but rather as part of the right way to handle and debate cultural assets. CIDOC-CRM[17] and FRBRoo[18] are the current state of the art to deal with complex, nuanced and opinionated annotations. Uncertainty needs to be assumed for every piece of information, and confidence and trust must be asserted based on agreement, reputation of the source and relevance of the assertion so as to distinguish undisputed vs. disputed vs. accepted facts. Geographical data is no exception to these problems, and is characterized by heterogeneity, differences of conceptual models, implementations, resolution and accuracy. Additionally, not all spatial relationships are explicitly represented, but inferred by algorithms over geometries. Geodata interoperability is sought by designing metadata standards and ontologies to align different conceptual models. This area will also consider workflows and models that respect the correct handling of digital rights, although this issue will be better addressed by Pillar 2.

Flexible Data Modeling for primary data includes work on different **Data Formats (1.2.2)**, which are strictly media-dependent, non-proprietary, feature rich and widely used, e.g., XML[19] for texts, TIFF[20] or JPG2000[21] for images, MXF D10[22] or H264[23] for video, MP3[24] for audio, STL[25] and COLLADA[26] for 3D scans are appropriate for primary data, while RDF is appropriate for secondary and tertiary data.

In the TM model every piece of information will be contextualized with a rich set of descriptors associated to multidimensional conceptual models. Time Machine will identify supplementary, rather than substitutive, metadata models to enrich to the ones already adopted locally, so as to guarantee continuity with the past but also homogenization and integration

---

[17] Le Boeuf D., and Ore S. (eds.), Definition of the CIDOC Conceptual Reference Model, volume A. In: *ICOM/CIDOC Documentation Standards Group*, 6.2.6, (May 2019).
http://www.cidoc-crm.org/sites/default/files/CIDOC%20CRM_v6.2.6_Definition_esIP.pdf, http://www.cidoc-crm.org/

[18] Bekiari, Doerr, Le Boeuf, and Riva (eds.), Definition of FRBROO, A Conceptual Model for Bibliographic Information in Object-Oriented Formalism. In: *Working Group on FRBR/CRM Dialogue, International Federation of Library Associations*, 2.4, (November 2015). https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf

[19] Bray, Paoli, Sperberg-McQueen, Maler, Yergeau (eds.), Extensible Markup Language (XML). In: *1.0 (Fifth Edition) W3C Recommendation*, (26 November 2008).
https://www.w3.org/TR/xml/

[20] Adobe Developers Association, TIFF Revision 6.0, Final — (June 3 1992).
https://www.adobe.io/content/dam/udp/en/open/standards/tiff/TIFF6.pdf

[21] ISO/IEC 15444-1:2019, JPEG 2000 image coding system, https://www.iso.org/standard/76647.html

[22] ST 377-1:2011 - SMPTE Standard - Material Exchange Format (MXF) — File Format Specification, (7 June 2011). doi: 10.5594/SMPTE.ST377-1.2011

[23] International Telecommunication Union (ITU-T). H.264, Advanced video coding for generic audiovisual services, 06/2019, https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-H.264-201906-I!!PDF-E&type=items

[24] ISO/IEC 13818-3:1998, Generic coding of moving pictures and associated audio information — Part 3: Audio.
https://www.iso.org/standard/26797.html

[25] Hiller, Lipson. STL 2.0: A proposal for a universal multi-material Additive Manufacturing File format. In: *20th Annual International Solid Freeform Fabrication Symposium*, pp. 266-278, SFF (2009).
https://sffsymposium.engr.utexas.edu/Manuscripts/2009/2009-23-Hiller.pdf

[26] ISO/PAS 17506:2012, COLLADA digital asset schema specification for 3D visualization of industrial data,
https://www.iso.org/standard/59902.html

between datasets. Such supplementary models will specifically address the characteristics of the metadata themselves rather than of the entities they describe, providing contextualization of these descriptors in terms of geographical and temporal aspects, provenance, authoritativeness, confidence, etc.

OWL-Time,[27] Prov-O,[28] the Confidence Information Ontology,[29] the Historical Context Ontology,[30] and the Time-Indexed Value in Context design pattern[31] are a good initial set of models. Due to the very different kind of CH objects to handle, we will develop appropriate **Metadata Formats and Mapping between Standards (1.2.3)**. TM will incorporate and link to already existing document and metadata models (e.g., DC,[32] CIDOC-CRM,[33] EAD,[34] METS,[35] LIDO,[36] TEI[37]) and to the vocabularies of values already in use (such as Geonames,[38] Getty Vocabularies,[39] Iconclass,[40] DBPedia,[41] etc.)

Metadata will be characterized in terms of their provenance, authoritativeness and acceptance, so that a variety of sources of annotations can be accepted even if of very different degrees of correctness; in particular, this allows low-cost, low-reliability automatic tools based on Artificial Intelligence, Machine Learning and Image Recognition algorithms to be used side by side with expensive, traditional, manual activities by domain experts, as long as the relative authoritativeness, reliability and richness is specified. TM will thus foster and support the creation of sophisticated algorithms for the automatic descriptions of cultural assets through Machine Learning and Artificial Intelligence algorithms, and will complement it with newer and easier interfaces to manual **annotation (1.2.4)** to offer to the LTMs. This interface will be created by following current knowledge of best practices in human-computer interaction and will use AI to

---

[27] Cox, Little (eds.), Time Ontology in OWL. In: *W3C Recommendation* (19 October 2017). https://www.w3.org/TR/owl-time/

[28] Lebo, Sahoo, McGuinness (eds.), PROV-O: The PROV Ontology. In: *W3C Recommendation* (30 April 2013). https://www.w3.org/TR/prov-o/

[29] Confidence Information Ontology, http://obofoundry.org/ontology/cio.html

[30] Daquino, Historical Context Ontology, http://hico.sourceforge.net/

[31] Peroni, Shotton, Vitali, Scholarly publishing and Linked Data: describing roles, statuses, temporal and contextual extents, In: *I-SEMANTICS '12: Proceedings of the 8th International Conference on Semantic Systems*, pp. 9-16 (September 2012). https://doi.org/10.1145/2362499.2362502

[32] DCMI Usage Board, Dublin Core Metadata Element Set, Version 1.1: Reference Description, (June 2012). https://www.dublincore.org/specifications/dublin-core/dces/

[33] Le Boeuf D., and Ore S. (eds.), Definition of the CIDOC Conceptual Reference Model, volume A. In: *ICOM/CIDOC Documentation Standards Group*, 6.2.6, (May 2019). http://www.cidoc-crm.org/sites/default/files/CIDOC%20CRM_v6.2.6_Definition_esIP.pdf, http://www.cidoc-crm.org/

[34] Technical Subcommittee for Encoded Archival Standards of the Society of American Archivists, EAD (Encoded Archival Description), 2012, https://www.loc.gov/ead/

[35] Network Development and MARC Standards Office of the Library of Congress, METS (Metadata Encoding and Transmission Standard) 1.12.1, 2019, https://www.loc.gov/standards/mets/

[36] Cobusn, Light, McKenna, Stein, Vitzthum, LIDO - Lightweight Information Describing Objects Version 1.0, November 2010, International Council of Museums, http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf

[37] Text Encoding Initiative (TEI), P5, November 2007, https://tei-c.org/guidelines/

[38] https://www.geonames.org/

[39] https://www.getty.edu/research/tools/vocabularies/

[40] http://www.iconclass.nl/home

[41] https://wiki.dbpedia.org/

assist in the annotation (cf. Task 2.4 in the taxonomy).

Regarding data modeling there are many relevant initiatives and projects on both European and international level that need to be considered (Europeana, Clarin, and others mentioned in the Stakeholders section). TM will not have to invent new models, but enforce models already used.

1.3. LONG-TERM PRESERVATION[42] describes the sum of processes undertaken to maintain a digital object's availability and interpretability – from a technological as well as a semantic perspective: digital preservation. Risks can be classified as falling into different classes. The **Bitstream Layer (1.3.1)** (Storage Layer) shall mitigate technical (e.g., bitflip, media failure), human (e.g., accidental deletion/manipulation) and environmental (e.g., fire) risks. Best practice for this relies on multiple independent copies of a digital object, stored in storage frameworks based on spinning disk and/or tape, often in hierarchical storage management systems. In recent years, the utilization of cloud storage providers has grown, and data integrity is verified via checksum algorithms which are typically stored in a local, centralised databases. There has also been some exploration of blockchain technology for a ledger containing integrity information. A potential game changer for the bitstream layer is DNA storage that is able to condense information on molecule level. The **Functional Layer (1.3.2)** shall mitigate risks associated with the technical interpretability of digital objects, mainly pertaining to the interpretation of the file formats in which the data information is encoded, as file formats may become technologically obsolete or may no longer fulfil user requirements.

The **Semantic Layer (1.3.3)** deals with risks associated with context or knowledge loss in the form of missing contextual information needed to interpret information within an object in the first place, or in the form of semantic change over the course of time. The TM should adhere to the requisites of **Trustworthy Archives (1.3.4),** proven through certification measures such as

[42] CCSDS, *Reference Model for an Open Archival Information System (OAIS)*, *Magenta Book*, CCSDS, (2012). http://public.ccsds.org/publications/archive/650x0m2.pdf;
CoreTrustSeal, *Core Trustworthy Data Repositories Requirements 2020 - 2022 Extended Guidance*, Version 2.0. https://www.coretrustseal.org/wp-content/uploads/2019/11/2019-10-CoreTrustSeal-Extended-Guidance-v2_0.pdf;
Dappert A., Guenther R., and Peyrard S., *Digital Preservation Metadata for Practitioners: Implementing Premis*, Springer International Publishing (2016).
DOI 10.1007/978-3-319-43763-7;
Digital Preservation Handbook, 2nd Edition, In: *Digital Preservation Coalition*. (2015).
http://handbook.dpconline.org/;
ISO. ISO 16363:2012 (CCSDS 652.0-R-1) Space data information transfer systems – Audit and certification of trustworthy digital repositories. (2012);
Lavoie, B. (2014). *The Open Archival Information System (OAIS) Reference Model: Introductory Guide*, 2nd Edition, In: *Digital Preservation Coalition* (2014).
https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file;
Lindlar M, Schwab F., All that work … for what? Return on Investment for Trustworthy Archive Certification Processes – A Case Study. In: *Proceedings of the 15th International Conference on Digtial Preservation*. iPRES 2018 (September 16-20 2018)
DOI 10.17605/OSF.IO/8A3SC;
Lindlar M., Saemann H., Ochmann S., Jonsson Ö., Gadiraju U., DURAARK Deliverable 6.6.1. Current State of 3D Object Digital Preservation and Gap Analysis Report, pp. 194. https://doi.org/10.5281/zenodo.1115503;
nestor Certification Working Group. Explanatory notes on the nestor Seal for Trustworthy Digital Archives. Nestor Materials (17 July 2013).
http://files.dnb.de/nestor/materialien/nestor_mat_17_eng.pdf;
Owens T., *Theory and Craft of Digital Preservation*. Johns Hopkins University Press, 1st Edition (December 11 2018).

CoreTrustSeal, DIN 31644 nestor Seal or ISO 16363 certification processes, and ISO 14721:2012 Open Archival Information System Standard, and be vigilant towards new developments, such as the forthcoming new ISO 14271:2019 version, now also including an outer-OAIS/inner-OAIS model, by means of which functional entities are spread across different organisations. The goal of the TM in this area is to create guidelines, best practices and recommendations, as well as a framework and a basic infrastructure that LTMs and other projects can apply to secure the preservation of their materials (to be part of the TM Recommendations). This should be a European and publicly guided infrastructure that preserves the continent's cultural heritage and secures the integrity of the data, working jointly and enhancing the already existing European initiatives.

Computing and Artificial Intelligence

Collecting and curating such a substantial amount of digital (meta)data will imply a significant contribution to Europe's engagement with its CH. Nevertheless, it only concerns one part of TM's ultimate objective. In order to deepen our insights in and understandings of Europe's past, present, and future, powerful technologies are needed that can bring the past back to life and re-invigorate our shared history. The TM will bootstrap from current trends in computing and AI and stimulate innovative research initiatives to improve the exploitation and understanding of CH objects. Especially in the field of Machine Learning, these will dramatically improve Europe's visibility and its contribution to global research challenges. The developments in these fields are essential for the creation of the TM Official Components. In this respect, Computing and AI are divided in the following areas:

2.1. COMPUTER VISION AND PATTERN RECOGNITION deals with ways in which computers can transform data into valuable information that can be processed and analysed. One of the most important subfields is **Text Recognition (2.1.1)**, seeing as a large part of the Big Data of the Past sources are texts in different forms: handwritten, printed, carved, etc. Currently, OCR systems show good accuracy if documents are clean and the type font is known,[43] and neural network models have shown success for handwritten texts in favourable situations.[44] However, for historical texts this is often not the case.[45] Therefore, TM has the potential to create an immense jump forward in this area: with enough data from different periods and places, there is a unique opportunity to train fully scalable, robust models for any kind of collection, as well as to generate high accuracy for heterogeneous documents. TM will design open access tools able to perform these tasks.

However, not all documents use text: some of them, such as maps, music scores and technical drawings, communicate information in their own specific languages and alphabets. The

[43] Zheng H., Chen K., Jianhua H., Xiang B., Karatzas D., Shijian L., Jawahar C.V, ICDAR2019 Competition on Scanned Receipt OCR. In: *Proceedings of the 15th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 1516-1520. (2019).

[44] Reeve Ingle R., Fujii Y., Deselaers T., Baccash J., Popat A. C., A Scalable Handwritten Text Recognition System. In: *Proceedings of the 15th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 17-24. (2019).

[45] Sánchez J. A., Romero V., Toselli A. H., Villegas M., Vidal E., ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset. *Proceedings of the 14th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 1383-1388. (2017).

field of **Graphic Document Processing (2.1.2)** will develop methods to process these kinds of documents. Printed mathematical and musical notations already demonstrate good results, but handwritten documents do not follow suit and require further research.[46] Furthermore, maps and architectural drawings stand out as giving rise to a difficult yet important problem which could be solved within the framework of the TM. Currently, there are only isolated, map-type oriented image processing algorithms,[47] as the uncertainty in the object representation has not yet been sufficiently studied and modelled. The goal of the TM in this area would be to create a common set of algorithms for processing old maps, and a European hub with standardised access for visualisation, download and annotation.

Image Processing and Analysis (2.1.3) is closely related to the two previous topics and researches general computer processing of old documents needed for many other tasks, such as enhancement and restoration, as well as segmentation and layout analysis. When framed as image-to-image translation tasks, in which both input and output are images, the current state of the art is Generative Adversarial Networks (GANs), still mainly applied to contemporary content.[48] These methods are restricted to translations for which there is (paired or unpaired) data from before and after the translation, or to domains for which we can artificially simulate the after situation (self-supervised learning). For layout analysis, there is good accuracy in standard layouts, but low performance in complex layouts (with multi-oriented texts, mixing of text with symbols, drawings, figures, strike-outs and hand-annotations etc.).[49] The objective of the TM in this area is to generate a rich framework which can perform realistic looking image-to-image translations between a wide-range of domains for many tasks, while incorporating side-information from historical sources. For example, possessing information about the function of a building and images from the neighbourhood can be an asset for creating a better reconstruction of the façade of a building.

Indexing and Retrieval (2.1.4) deals with methods that make searching the content efficient and successful. Given the amount of data TM will generate, this is a crucial area.

---

[46] Pacha A., Choi K.-Y., Coüasnon B., Ricquebourg Y., Zanibbi R., Eidenberger H., Handwritten music object detection: Open issues and baseline results. In: 1*3th IAPR International Workshop on Document Analysis Systems* (DAS), pp. 164-168. (2018);
Baró A., Riba P., Calvo-Zaragoza J., Fornés A., From optical music recognition to handwritten music recognition: A baseline. In: *Pattern Recognition Letters* 123, pp. 1-8. (2019);
Mahdavi M., Zanibbi R., Mouchere H., Garain U., ICDAR 2019 CROHME + TFD: Competition on Recognition of Handwritten Mathematical Expressions and Typeset Formula Detection. In: *Proceedings of the 15th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 1533-1538. (2019).
[47] Weinman J., Chen Z., Gafford B., Gifford N., Lamsal A., and Niehus-Staab L., Deep Neural Networks for Text Detection and Recognition in Historical Maps. In: *Proceedings of the 15th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 902-909. (2019).
[48] Yeh R. A., Chen C., Lim T. Y., Schwing A. G., Hasegawa-Johnson M., Do M. N., Semantic image inpainting with deep generative models. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5485-5493. (2017);
Nguyen K. C., Nguyen C. T., Hotta S., Nakagawa M., A Character Attention Generative Adversarial Network for Degraded Historical Document Restoration. In: *Proceedings of the 15th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 420-425. (2019).
[49] Clausner C., Antonacopoulos A., Pletschacher S., ICDAR2019 Competition on Recognition of Documents with Complex Layouts - RDCL2019. In: *Proceedings of the 15th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 1521-1526. (2019);
Gao L., Huang Y., Déjean H., Meunier J.-L., Yan Q., Fang Y., Kleber F., Lang E., ICDAR 2019 Competition on Table Detection and Recognition (cTDaR). In: *Proceedings of the 15th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 1510-1515. (2019).

Currently, the cutting-edge solution is Probabilistic Indexing,[50] where the user can choose between more precision (which entails missing true positives) or recall (which entails generating false positives). Nowadays, Probabilistic Indexing on collections composed with hundreds of thousands of handwritten documents are available as prototypes. On the other hand, content-based image indexing and retrieval (CBIR) tools provide efficient performances with contents of the same domain, but when considering cross-domain images indexing and retrieval perform poorly. The goal of the TM is to develop, as part of the TM Components, uni-modal and cross-modal repeating motif discovery on a super-large scale, as well as search engines that are completely operative based on probabilistic indexing and capable of operating for any kind of cross-domain content.

The field of **Understanding and Interpretation (2.1.5)** enables reasoning with reference to visual content,[51] video and image captioning,[52] and the creation of tools to automatically gain information about semantics, style, dating, and location of a document.[53] This is a domain in which major breakthroughs should be expected in the upcoming years, but not necessarily with a cultural heritage focus. The TM should, on the one hand, redirect such new technologies to this field and, on the other, actively contribute to the area by taking advantage of the materials collected and by means of the expertise of its members. Being able to infer geographical information will be a particularly strong TM component to improve simulation engines.

Recognition and Detection (2.1.6) focuses on the problem of identifying entities (persons, animals or objects) in different kinds of images (photographs, paintings, films, videos, 3D). Currently, a good performance is achievable for data-rich, single-modality recognition and detection of prominent classes for contemporary content with deep learning-based techniques.[54] The aim of the TM is to adapt and improve these technologies for the Big Data of the Past which entails a specific set of problems due to the particularities of the material.

A particularly pertinent problem of recognition and detection is **Person and Face**

[50] Toselli A. H., Romero V., Vidal E., Sánchez J. A., Making Two Vast Historical Manuscript Collections Searchable and Extracting Meaningful Textual Features Through Large-Scale Probabilistic Indexing. In: *Proceedings of the 15th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 108-113. (2019).
[51] Santoro A., Raposo D., Barrett D. G., Malinowski M., Pascanu R., Battaglia P., Lillicrap T., A simple neural network module for relational reasoning. In: *Advances in neural information processing systems*. pp. 4967-4976. (2017).
[52] You Q., Jin H., Wang Z., Fang C., Luo J., Image captioning with semantic attention. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4651-4659. (2016).
[53] Wahlberg F., Wilkinson T., Brun A., Historical manuscript production date estimation using deep convolutional neural networks. In: *15th International Conference on Frontiers in Handwriting Recognition* (ICFHR). IEEE. pp. 205-210. (2016);
Gao L., Huang Y., Déjean H., Meunier J.-L., Yan Q., Fang Y., Kleber F., Lang E., ICDAR 2019 Competition on Table Detection and Recognition (cTDaR). In: *Proceedings of the 15th International Conference on Document Analysis and Recognition* (ICDAR), 1, pp. 1510-1515. (2019);
He S., Samara P., Burgers J., Schomaker L., Historical manuscript dating based on temporal pattern codebook. In: *Computer Vision and Image Understanding*, 152, pp. 167-175. (2016).
[54] He K., Gkioxari G., Dollár P., Girshick R., Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2961-2969. (2017);
Redmon J., Farhadi A., Yolov 3: An incremental improvement. (2018). arXiv preprint arXiv:1804.02767

Identification (2.1.7). Issues revolving around gender,[55] race[56] and age[57] bias in the training materials of current models is an important issue to address,[58] as well as the need to develop models that can work with limited training data from historical sources, which are derived from heterogenous objects including statues, paintings or coins.

Modelling, Registration, and Reconstruction (2.1.8) entails using geographical data (maps, satellite images, geo-referencing) to create digital models and reconstructions of historical sites. Currently, there is no uniform processing pipeline and only local, site-specific long-term reconstruction samples are available.[59] Automated tools for national mapping agencies exist for the generalisation of topographic data and maps, but these do not often allow for real-time operation. Currently, automatic georeferencing and spatial conflation only work for large map series (topographic map series). The goal of the TM will be to create a European hub for geodata with standardised access for visualisation; a common CS platform for different tasks (georeferencing, digitisation, training sample collection); and a cloud computing (HPC/storage) infrastructure for storing and processing maps, remote sensing imagery and the derived reconstruction results (including a versioning system).

Finally, the TM will also work with **Audio Recognition and Transcription (2.1.9)**. The TM can aim at creating universal models that allow the transcription of historical audio and audio-visual contents in most of the languages of the European Union and its neighbouring territories.[60]

[55] Muthukumar, V., Color-Theoretic Experiments to Understand Unequal Gender Classification Accuracy from Face Images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. (2019).
[56] Vangara K., King M. C., Albiero V., Bowyer K., Characterizing the Variability in Face Recognition Accuracy Relative to Race. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. (2019).
[57] Srinivas N., Ricanek K., Michalski D., Bolme D. S., King M., Face Recognition Algorithm Bias: Performance Differences on Images of Children and Adults. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. (2010).
[58] Kortylewski A., Egger B., Schneider A., Gerig T., Morel-Forster A., Vetter T., Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 1093-2102. (2018).
[59] Kaplan, F., The Venice Time Machine. In: *Proceedings of the 2015 ACM Symposium on Document Engineering*. (2015).
[60] Nouza J., Cerva P., Zdansky J., Blavka K., Bohac M., Silovsky J., Chaloupka J., Kucharova M., Seps L., Malek J., Rott M., Speech-to-text technology to transcribe and disclose 100,000+ hours of bilingual documents from historical Czech and Czechoslovak radio archive. In: 1*5th Annual Conference of the International Speech Communication Association* (INTERSPEECH). (2014);
Boyd D., OHMS: Enhancing access to oral history for free. In: *The Oral History Review*, 40, pp. 95-106. (2013).

2.2. NATURAL LANGUAGE PROCESSING (NLP)[61] is of utmost importance for the TM: many of the documents to be digitised include text in a natural language. This field includes some of the most important and revolutionary innovations as machines get exponentially better at understanding human language. However, most of those innovations are aimed primarily at the leading modern languages, in particular English. The TM will facilitate the usage of these revolutionary technologies for contemporary as well as older language variants for as many European languages as possible.

Multilinguality is now better handled thanks to multilingual language models encoding information simultaneously for different languages. These models also make it possible to develop language transfer techniques, with two main benefits: i) because these models better handle language variation, NLP processes are more robust across text genres and domains and ii) it is

---

[61] Adams O., Makarucha A., Neubig G., Bird S., and Cohn T., Cross-lingual word embeddings for low-resource language modeling. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 1, pp. 937-947. (2017);
Alex B., Degaetano-Ortlieb S., Kazantseva A., Reiter N., and Szpakowicz S., Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. (2019);
Brando, C., and Frontini F., Semantic Historical Gazetteers and Related NLP and Corpus Linguistics Applications. pp. 1-6. (2017);
Casey, A., Bennett M., Tobin R., Grover C., Engelmann L., and Alex B., Plague Dot Text: Text mining and annotation of outbreak reports of the Third Plague Pandemic (1894-1952). (2019);
Cherry C., Durrett G., Foster G., Haffari R., Khadivi S., Peng N., Ren X., and Swayamdipta S., Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP* (2019);
Jurafsky D. and Martin J.H., Speech and Language Processing (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/ (1st and 2nd ed. Prentice Hall). (2019);
Liu B., Sentiment Analysis and Opinion Mining. San Raphael: Morgan & Claypool (Synthesis Lectures on Human Language Technologies). (2012);
Nie J.Y. Cross-Language Information Retrieval. San Raphael: Morgan & Claypool (Synthesis Lectures on Human Language Technologies). (2010);
Palmer M., Gildea D., Xue N., Semantic Role Labeling. San Raphael: Morgan & Claypool (Synthesis Lectures on Human Language Technologies). (2010);
Pettersson E., Megyesi B., and Nivre, J., Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In: *Proceedings of the 19th Nordic Conference of Computational Linguistics* (NODALIDA 2013), *NEALT Proceedings Series*, 16.085, pp. 163-179. Linköping University Electronic Press. (2013);
Piotrowski M., Natural Language Processing for Historical Texts. San Raphael: Morgan & Claypool (Synthesis Lectures on Human Language Technologies). (2012);
Poibeau T. *Machine Translation*. Cambridge: MIT Press. (2017);
Poibeau T., Saggion H., Piskorski J., Yangarber R., eds., *Multi-source, Multilingual Information Extraction and Summarization*. Heidelberg: Springer. (2013);
Reynaert M., Hendrickx I., and Marquilhas R., Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2. In: *Annotation of Corpora for Research in the Humanities* (ACRH-2), pp. 87. (2012);
Saggion H., Automatic Text Simplification. San Raphael: Morgan & Claypool (Synthesis Lectures on Human Language Technologies). (2017);
Søgaard A., Vulić I., Ruder S., Faruqui M. Cross-Lingual Word Embeddings. San Raphael: Morgan & Claypool (Synthesis Lectures on Human Language Technologies). (2019);
Sprugnoli R., and Tonelli S., Novel event detection and classification for historical texts. In: *Computational Linguistics*, 45.2, pp. 229-265. (2019);
Wevers M., Hasanuzzaman M., Dias G., Düring M., Jatowt A., Proceedings of the the 5th International Workshop on Computational History, *HistoInformatics* (2019);
Yadav V., Bethard S., A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. Proc of Coling. (2018).

now possible to develop effective tools for low resource languages, even without large parallel corpora. Multilingual models can be used for a large variety of tasks, from named entity recognition to parsing and semantic reading.

At the same time, as a LSRI the TM could provide a platform for research teams to share their resources, support research initiatives and become a hub of developments. Of course, this will be done in collaboration with existing research infrastructures such as CLARIN and DARIAH – not competing with them, but using the existing infrastructures to make the TM's output available via sustainable, already publicly funded existing initiatives, which will in turn be enhanced by the TM's contribution.

Main tasks in this domain are **Methods for Resource-Scarce Languages (2.2.1)**, which aim to get language technology up to speed for different languages. To achieve this, one of the strategies is transfer learning and attention modelling, to transfer knowledge and models from one language to another and to bootstrap language technology tools for languages that are less wide-spread.

Another important issue for the kind of data the TM will be handling is **Orthographic Normalisation and Variation Handling (2.2.2)**, as older documents tend to be written with other orthographic conventions than modern ones and many NLP tools are optimised to contemporary spelling and text conventions.

The third core topic in this domain is **Machine Reading / Document Understanding / Question Answering (2.2.3)** which involves the ability of a computer to find relevant information in a text. The main issue in this task is text understanding, or the mapping of natural-language sentences to formal representations of their meaning. The current state of the art is not robust against text and genre variations. Two huge leaps forward would be to go beyond the sentence and document boundaries and achieve true multilingual support. Another relevant task in this area is information extraction, which includes Named Entity Recognition (NER), disambiguation and linking, where, again, the focus will be in creating resources for smaller European languages. In order to ensure that the results of the normalisation and machine reading tasks are stored properly and connected to other information about the documents, the task **Metadata Extraction, Manipulation, and Translation / Mapping (2.2.4)** has been devised. The research of this task focuses on designing and improving methods to extract metadata automatically from text documents and to map existing metadata into a standard metadata format. This task force will work closely with the team in charge of task 1.2 "Data Modelling". **Discourse analysis (2.2.5)** is concerned with uncovering the author's intent and will create tools to mine sentiments and opinions from texts as well as analyse trends.

In all NLP tasks, the type of data and the time of composition are the main issues to tackle, as most available tools were designed for modern language variants. The TM will expand this current scope to the Big Data of the Past. Machine translation could prove to be a very important aspect of the TM, as it deals with many language variants. Given the fact that TM is a European project, it should be available for speakers of many different languages. State-of-the-art technology in this task involves machine learning (cf. task 2.3.). TM should not only be able to offer reliable translations of older language variants into the corresponding modern languages, but also to other European languages.

2.3. MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE[62] is an autonomous field of development for the project and this transversal discipline will have repercussions on many other areas. **General Artificial Intelligence (2.3.1)** aims at solving tasks that are usually considered to require human intelligence. Strong AI that is able to perform human-level reasoning is as yet unknown. TM might bring significant advances in this field due to the large amount and variety of annotated data that will become available thanks to the project. One promising first step in this direction is meta learning that aims at the learning process of model discovery and re-use. Furthermore, the creation of causal models is a promising approach towards the creation of interpretable and re-useable models.

Currently, most of Artificial Intelligence is done through Machine Learning, which develops methods for learning parameters of computational models from sampled data. Three major directions towards this goal can be distinguished and they are considered separately in the TM's taxonomy. In **Supervised Learning (2.3.2)** a large, representative data set is available that is completely annotated with the desired meta-information that the trained model should produce for new, unseen samples after having been trained. In contrast to this, for **Unsupervised Learning (2.3.3)** only the data set but no label information is available. Consequently, unsupervised learning can identify hidden structure in data. In **Weakly Supervised Learning (2.3.4),** situations are considered in which annotated data is scarce or not available for the target domain at all, for example when only a small fraction of the data set available for training is annotated, or when no labels are given but the learning algorithm is allowed to query some labels for selected samples. In **Transfer Learning (2.3.5)** an annotated data set is only available for a problem similar to the one considered but not for the target domain itself. Consequently, models learned on some other set of samples are transferred and applied to a related domain. All these cases are especially relevant for the TM, as the data from cultural heritage objects is heterogenous and the quality of the existent annotation may vary widely.

62

Bishop C. M., *Pattern recognition and machine learning.* Springer (2007);

Duda R. O., Hart P. E., and Stork D. G., *Pattern classification.* Wiley (2001);

Hastie T., Tibshirani R., and Friedman J. H., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer (2009);

Murphy K. P., *Machine learning - a probabilistic perspective.* The MIT Press (2012);

Chapelle O., Schölkopf B., and Zien A., *Semi-Supervised Learning.* The MIT Press (2006);

Zhu X., and Goldberg A. B., *Introduction to Semi-Supervised Learning.* Morgan & Claypool Publishers (2009);

Goodfellow I., Bengio Y., and Courville A., *Deep learning.* The MIT Press (2016);

LeCun Y., Bengio Y., and Hinton G. E., Deep Learning. In: *Nature*, 521.7553, pp. 436-444. (2015);

Liu C.-L., Fink G. A., Govindaraju V., and Jin L., Special issue on deep learning for document analysis and recognition. In: *International Journal of Document Analysis and Recognition* (IJDAR), 21.3, pp. 159-160. (2018);

Hohman F., Kahng M., Pienta R. and Chau D. H., Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. IEEE Trans. In: *Vis. Comput. Graph*., 25.8, pp. 2674-2698. (2019);

Samek W., Montavon G., Vedaldi A., Hansen L. K., and Müller K.-R., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. 11700. Springer (2019);

Zhang Q., and Zhu S.-C., Visual interpretability for deep learning: a survey, In: *Frontiers of IT & EE*, 19. 1, pp. 27-39. (2018);

Kamishima T., Akaho S., and Sakuma J., Fairness-aware Learning through Regularization Approach. In: *Data Mining Workshops* (ICDMW) (2011).

Many of the recent breakthroughs in Machine Learning have been made using artificial neural networks, leading to the emergence of a new area called **Deep Learning (2.3.6)**. Models in this area set themselves apart from classical models by considerably ramping up model complexity. Incorporating and contributing developments in this area into the TM will accelerate the ability to analyse and understand the Big Data of the Past. In particular, research towards a **Universal Representation Space (2.3.7)** that is not only able to describe meaning and semantics of objects, text, images, and other information sources, but that is also able to transfer all of these representations into each other by means of recent progress in machine translation, is a major aim for TM.

For any machine learning framework or method one can ask meta-questions. In this regard, **Explainability (2.3.8)** of models is a desideratum if predictions are to be accepted by domain experts. Understanding the failure of a model and identifying how to address these cases also form issues that should be addressed. In this regard, fusion with prior knowledge is a promising technique towards this goal. While all data-driven approaches rely on training data, the open question remains of how to minimize unwanted implications of the composition, selection and availability of training material in order to allow algorithms and applications to be fair and unbiased. This should be addressed as part of the topic **Bias and Fairness (2.3.9**) which is extremely relevant for cultural heritage objects that were produced and preserved in particular societies for which it is necessary to be aware of the biases they transport to learning models.

2.4. HUMAN-COMPUTER INTERACTION AND VISUALISATION considers how the human end-user will interact with the materials generated in the project, from accessing the data and metadata to asking research questions. The main task here revolves around **User-centred interfaces (2.4.1)**, which deal with creation models and interfaces for end-users. These interfaces will be fundamental for the Targeted Achievements 2.1 and 2.2 (see below). The TM needs to create interfaces for annotation and for end-users that are flexible and easy to understand. Most of the other fields in this area relate to the end-user interface, which enables the interaction with the Big Data of the Past. **Access to large-scale information retrieval and recommender systems (2.4.2)[63]** deals with the ways in which user queries will be processed for such a huge database and how the information of searches will be harvested by the project to improve functionality. This includes browsing, exploratory search, intelligent user interface, and large-scale knowledge retrieval. Currently, many sources can be available to index an image document (image content, metadata, reference frames, geometry of the scene, etc.) but they perform poorly when used jointly. Deep learning techniques have just begun to manage them jointly with a mitigated efficiency, for example by generating missing modalities in the input when querying a multimodal dataset.

In the field of **virtual, augmented, and mixed reality (2.4.3)[64]**, there are currently multiple

[63] Mai J. E., Looking for information: A survey of research on information seeking, needs, and behavior. Emerald Group Publishing (2016).

[64] Bekele M. K., et al. A survey of augmented, virtual, and mixed reality for cultural heritage. In: *Journal on Computing and Cultural Heritage* (JOCCH), 11.2, pp. 1-36. (2018);
Zak E., Do you believe in magic? Exploring the conceptualization of augmented reality and its implications for the user in the field of library and information science. In: *Information Technology and Libraries*, 33.4, pp. 23-50. (2014);
Banfi F., and Oreni D., Virtual Reality (VR), Augmented Reality (AR), and Historic Building Information Modeling (HBIM) for Built Heritage Enhancement: From Geometric Primitives to the Storytelling of a Complex Building. In:

technologies (HMDs, CAVEs, Powerwalls, L-shapes, etc.) that are already used partially in classrooms, museums, and games. There are however, many challenges in user positioning, rendering and visualisation design. The TM will contribute to that development by focusing on how 3D and 4D models of cultural heritage objects, buildings and sites can be incorporated in virtual, augmented, or mixed reality environments for GLAM, education, and games.

Accessibility and learning, adaptive and cognitive interfaces (2.4.4)[65] deal with user interfaces that adapt as best as possible to the needs of users and that are optimised for learning experiences. This area includes, for example, remote teaching using combined video-conferencing, 3D virtual models and motion tracking, fully distributed learning environments or natural user interfaces to interact collaboratively with remotely rendered 3D models. In its efforts to actively construct the future using the Big Data of the Past, the TM should guarantee that users have the best user interfaces possible and innovate in the dissemination and teaching thereof.

Motivational design (2.4.5)[66] (gamification, storytelling) is a design strategy to increase motivation and participation, in other words, to motivate visitors to interact with places of interest and to support a deeper immersion. Elements such as badges or rankings can evoke strong positive as well as negative emotions in the target group, whereas competition-promoting approaches can be experienced as motivating and challenging: the attitudes and expectations of the target group should always be inquired about in advance when selecting gamification elements. **Big data visualisation (2.4.6)** deals with specific problems of transforming big amounts of data into graphs, plots, maps and other kinds of visualisations that enable the human observer to understand patterns and to draw reasonable conclusions. It includes real-time in-situ visualisation, which demands an efficient development of new methods. Seeing as progressive data analysis is still rudimentary while the data volume keeps increasing, the TM will improve technology to create real time visualisations of 3D objects, buildings and cities, devising ways of efficient metadata representation.

The topic **User Experience (2.4.7)**[67] engages human perception and ergonomics in an attempt to improve the interaction between users and interfaces. Design patterns are widely available for 2D content, yet limited for 3D content. There are various visual design strategies such as Skeuomorphism – the imitation of real-life, familiar objects – yet the presentation on one single page, and bold and graphic/photo dominated typos need to be considered and improved.

Impact of Industry 4.0 on Architecture and Cultural Heritage. IGI Global, pp. 111-136. (2020).

[65] Peissner M., Häbe D., Janssen D., and Sellner T., MyUI: generating accessible user interfaces from multimodal design patterns. In: *Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems*, pp. 81-90. (June 2012).

[66] Peng A. C., Van Dyne L., and Oh K., The influence of motivational cultural intelligence on cultural effectiveness based on study abroad: The moderating role of participant's cultural identity. In: *Journal of Management Education*, 39.5, pp. 572-596. (2015).

[67] Koutsabasis P., Empirical evaluations of interactive systems in cultural heritage: a review. In: International Journal of Computational Methods in Heritage Science (IJCMHS), 1.1, pp. 100-122. (2017);
Konstantakis M., Michalakis K., Aliprantis J., Kalatha E., and Caridakis G., Formalising and evaluating cultural user experience. In: *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization* (SMAP), IEEE, pp. 90-94. (July 2017);
Piccialli F., and Chianese A., Cultural heritage and new technologies: Trends and challenges. In: *Virtual Research Environment*, 2.4.8. (2017);
Jouan P. A., and Hallot P., Digital Twin: a Hbim-Based Methodology to Support Preventive Conservation of Historic Assets Through Heritage Significance Awareness. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, pp. 609-615. (2019).

Regarding testing different kinds of user experiences, small user studies (eye tracking, observation, sound, movement, gesture) and data analysis (customer journeys, click paths) are common. Ergonomics in terms of device- and application-design is well researched, but very much depends on the device or OS. Current research involves the two key modes of media perception – sight and hearing – as well as kinaesthetic intelligence (movement, balance, coordination). In the context of Virtual and Augmented Reality, perception is very much researched in terms of avoiding biases by physiological effects such as motion sickness or plausibility.

Virtual Research Environment (2.4.8) is based on digital twins, a digital representation that mirrors a real-life object, process or system. These are robust models, linked to the real world and driven by AI, which enable interaction in "what-if" scenarios (cf. task 3.1.3 "Simulation"). The TM will investigate the creation of digital twins not only of objects, but more importantly of complex scenarios, like past cities and communities.

2.5. COMPUTER GRAPHICS create the necessary technology to display faithful and high-quality images which are very important for the TM: faithful representations of objects and spaces are required to engage scholarship and the general public. The creation of computational images of historical places and objects has been widely researched in the past, and the TM has the opportunity to create great improvements in this area thanks to the availability of data and a network of computing experts.

Rendering (2.5.1) is the automatic process of generating an image from a 2D or 3D model. Currently, high quality offline rendering is available to reconstruct the impression of artefacts in their original environment, and real time rendering of reconstructed sites is available in Virtual Reality (VR) or Augmented Reality (AR). The TM will improve high-quality rendering of historic artefacts in the context of their original environment, both in VR and AR.

Animation (2.5.2) covers the ability to create virtual images in movement. Currently, motion capturing using specialized suits and fine facial capture by means of a large set of facial markers are ideal processes to create models for individuals. At the same time, AI can generate convincible character animation and crowds.

Immersive, Virtual and Augmented Reality (2.5.3) is excellent in allowing exploration and interaction with historical reconstructions of the past, yet this is mainly associated with Head Mounted Displays (HMD) at present. Spatial AR is an emerging topic, which uses projectors. The TM will track the developments in the field and create appropriate technology for historical reconstructions: involving real-time tracking of complex scenes, and real-time rendering of complex scenes interacting with the real-world, based on a real-time 3D-reconstruction of the environment.

Interactive Computer Graphics and Computer Games (2.5.4) study ways to engage users with computer-generated images by giving them agency, the technological counterpart of topic 2.4.5. "Motivational Design". Another related problem is **Procedural Content Generation (2.5.6)**, also known as open world generation, which is especially relevant for big simulations of historical cities. Currently, shape Grammars/Procedural Descriptions are used. The TM will apply these technologies, focusing on CH, in order to derive generalised rules, that can generate objects for a certain style (i.e. gothic churches, 18th-century ships, etc.).

2.6. SUPER COMPUTING investigates the best way to fulfil the complex computing capacities for data acquisition, storage, processing and analysis needed at different stages of the TM. The TM must be aware of the existing technologies and infrastructures and develop methodologies if necessary.

Scaling and distribution (2.6.1) will keep track of important developments on large-scale and distributed computers. At present, there are several efforts to reach ExaScale Computing. From the hardware perspective, the problems are mainly the physical limits and growing power requirements for supercomputers that slow down the goal to build a first supercomputer. Furthermore, these supercomputers become more complex and often consist of heterogeneous architectures. A software redesign is thus required in order to adapt to the new hardware. Although Exascale computing will probably not entail general purpose machines and will not be available for the TM project, the architecture of the Exascale machines can provide a good example of technologies offered by future computing centres.

Dynamic provision of computing platforms (2.6.2) deals with the more concrete problem of creating computing platforms for the TM. Computing centres should set up computing platforms on demand to support requirements of the wide range of applications. The TM should also try to achieve a standardised API. This is related to the shared computing infrastructure described in more detail in the Roadmap for Pillar 2, as by the next two fields, namely **Cloud computing (2.6.3) which** investigates if this widely used technology could be applied for the TM, considering its advantages and potential problems, and **Secure distributed computing (2.6.4) which** is in charge of defining the requirements so that the TM computing infrastructure can perform safely and efficiently in a distributed way.

Social Sciences and Humanities

Time Machine has the capacity to create revolutionary approaches in the Humanities and Social Sciences, but it is necessary to engage researchers in these areas to create such new models that harvest the power of the Big Data of the Past. By having access to these data and applying the relevant methods it will be possible, for example, to plan the future development of cities considering their past; to better assess the cultural effects of climate change in past centuries; to explore precise traces of multiple migratory, commercial and artistic movements, consigned to the archives, bringing them back to life in the form of a great modelling of European circulation.

The specific contribution of the Social Sciences and Humanities area within pillar one of the Time Machine project lies in establishing a vivid arena for exchange between computer science and the humanities to stimulate a mutual process of exchange between the domains. Social Sciences and Humanities are not just consumers of technology but will provide significant impact on the design of systems by providing feedback on the use of new tools and toolchains and by critically reflecting and thereby challenging the conceptual models underlying the developments in the other areas.

Our approach to enable this process is two-fold: on the one hand the new technologies developed in other areas of pillar one will provide a strong technology induced push that will stimulate reflection in the different domains of humanities leading to new approaches and research questions. At the same time, we will use concrete research challenges from within the (digital) humanities community to create a demand (pull) for specific technological solutions with

the idea to combine the two effects to create a variety of different research projects.
Based on the experiences gathered from the different pilot research projects we will build platforms that enable the use of cutting-edge technologies to a broad range of different users that will in turn create significant interest in the humanities community thereby strongly shaping the acceptance of the digital tools and approaches developed within the Time Machine project.

The TM's strategy in this area is divided in two areas fields: "Theory" and "Disciplines".

3.1. THEORY deals with general questions regarding the relationship between TM and Social Sciences and Humanities (SSH), as well as considering the particular ways in which the project relates to society at large. The field **Qualitative and Quantitative SSH (3.1.1)** [68, 69, 70, 71, 72] will foster a dialogue between quantitative and qualitative approaches in historical research. The aim is to provide a conceptual and methodological framework for SSH scholarship which can combine the strengths of hermeneutic research (interpreting the complexity of human culture and society at the micro level of close-reading individual sources, places, people or events) with the

---

[68] Bergman M. M., The straw men of the qualitative-quantitative divide and their influence on mixed methods research. M.M. Bergman (ed.), *Advances in mixed methods research.* pp. 10-21. (2008);
Bryman A., Integrating quantitative and qualitative research: how is it done? In: *Qualitative Research*, 6.1, pp. 97-113. (2006).
Maxwell J. A., Expanding the History and Range of Mixed Methods Research. In: *Journal of Mixed Methods Research*, 10.1, pp. 12-27. (2016).
[69] Aiden, E., and Michel, J.-B., Uncharted: Big Data as a Lens on Human Culture. New York: Riverhead Books (2013);
Benardou A., Constantopoulos P., Dallas C., and Gavrilis D, Understanding the Information Requirements of Arts and Humanities Scholarship. In: *International Journal of Digital Curation*, 5.1, pp. 18-33. (22. June 2010). https://doi.org/10.2218/ijdc.v5i1.141;
Burnard L., Science and the Humanities. *Foxglove* (blog), 16, (March 2014). http://foxglove.hypotheses.org/468;
Epstein J. M., Why Model? In: *Journal of Artificial Societies and Social Simulation*, 11.4, (2008). http://jasss.soc.surrey.ac.uk/11/4/12.html
Gradmann S., Vom Verfertigen der Gedanken im digitalen Diskurs: Versuch einer wechselseitigen Bestimmung hermeneutisch und empirizistischer Positionen. *Historical Social Research*, 29.1, pp. 56–63. (2004). http://www.ssoar.info/ssoar/handle/document/3086
Han J., Data mining: concepts and techniques. 3rd ed. Burlington, MA: Elsevier (2011). http://web.engr.illinois.edu/~hanj/bk2/toc.pdf;
Lemercier C., Zalc C., and Goldhammer A., *Quantitative Methods in the Humanities: An Introduction*. Charlottesville; London: University of Virginia Press. Retrieved February 4, 2020, (2019). www.jstor.org/stable/j.ctvbqs963;
Svensson P., and Goldberg, D. T. eds., Between humanities and the digital. Cambridge: MIT Press (2015). http://ieeexplore.ieee.org/servlet/opac?bknumber=7120877 [Accessed May 19, 2019]
[70] https://foundhistory.org/2010/05/wheres-the-beef-does-digital-humanities-have-to-answer-questions/
Thiel T., Eine empirische Wende für die Geisteswissenschaften? *Frankfurter Allgemeine Zeitung* (FAZ) (2012). https://www.faz.net/aktuell/feuilleton/forschung-und-lehre/digital-humanities-eine-empirische-wende-fuer-die-geisteswissenschaften-11830514.html
[71] Liu A., Where Is Cultural Criticism in the Digital Humanities? In: *Debates in the Digital Humanities*, ed. M. K. Gold (U of Minnesota P) (2016). http://dhdebates.gc.cuny.edu/debates/text/20;
Ginzburg C., Signes, traces, pistes: Racines d'un paradigme de l'indice. In: *Le Débat*, 6.6, pp. 3. (1980). doi:10.3917/deba.006.0003;
Mühlberger G., Die digitale Wissenschaft. Einige Bemerkungen über das Verhältnis von Internet und Geisteswissenschaft. In: *Sprachkunst* 30.1, pp. 135-58. (1999). http://www.literature.at/viewer.alo?viewmode=overview&objid=21609&setLang=de
Presner T., Critical Theory and the Mangle of Digital Humanities. In: *The Humanities and the Digital*, ed. David Theo Goldberg and Patrik Svensson, MIT Press, pp. 20. (2012). http://www.toddpresner.com/wp-content/uploads/2012/09/Presner_2012_DH_FINAL.pdf
Scheinfeldt T., Where's the Beef? Does Digital Humanities Have to Answer Questions? In: *Found History* (2010).
[72] Fitzpatrick K., Planned obsolescence: publishing, technology, and the future of the academy. New York: New York University Press (2011).

advantages of quantitative methods (seeing patterns in large datasets and analysing those with statistical methods). Since the 1950s, the use of quantitative methods in historical research has waxed and waned. In recent years, newer subdisciplines like cultural history have largely ignored or even rejected quantitative analysis as they posed questions, even as the potential for application has risen through the exponential growth of digitized and born-digital sources. Today's quantitative methods include machine learning and probabilistic models that are well-suited to these more "qualitative" historical questions, unlike the more rigid statistical methods of the twentieth century. The TM will try to exploit this new environment, by showing how large scale historical data can enrich history-related research, establishing best practices for scholars about digitized and born-digital sources, and emphasizing the capacity of the humanities to participate in interdisciplinary projects.

Increased research scope in SSH (3.1.2)[73] focuses specifically in the ways in which the access to the Big Data of the Past can improve research in SSH. Traditional disciplines such as history and philology have primarily worked with a methodology that induces general principles based on a limited number of sources and observations. The TM offers the possibility to exponentially expand the sources available. However, those disciplines still need to extend their existing strategies to handle this new availability of digitised materials.[74] The TM needs to actively contribute to the creation of this paradigm shift, and not only by providing data. Therefore, in collaboration with Pillar 3 (Task 3.1 Scholarship) we will provide use cases on a variety of topics that bring a longitudinal perspective to present-day problems discussed in the domains of history and philology. These will yield best practices of what can be achieved with Big Historical Data and Scalable Humanities, not only in terms of knowledge but also in terms of methodology, and will be accompanied by training and dissemination materials that will be made available through our collaboration with the relevant domain organisations (via workshops and papers at their annual conferences) and European Research Infrastructures (CLARIN, DARIAH, EHRI, E-HRIS). The goal is to provide a situation that enables scholars to conduct research on their traditional research topics more easily than ever, choosing their scale of reading, playing with multiscale readings, crossing different kinds of sources, choosing the timespan of their research, being able

---

[73] Allen L., Introducing the Computing Cultural Heritage in the Cloud Project [Blogpost]. In: *The Signal*, Library of Congress Blog, November 25, (2019) https://blogs.loc.gov/thesignal/2019/11/introducing-the-computing-cultural-heritage-in-the-cloud-project/;
https://labs.loc.gov/experiments/cchc/?loclr=twndi;
http://en.statsbiblioteket.dk/kulturarvscluster/;
Blanke T., Predicting the Past. In: *Digital Humanities Quarterly*, 12.2, (2018);
Borgman C. L. *Big data, little data, no data: Scholarship in the networked world*. MIT press. (2015);
Hawkins R. A., The use of big data in historical research. In: Schiuma G., and Carlucci D. (eds.). *Big Data in the Arts and Humanities. Theory and Practice* (2018);
Jockers M. L., *Macroanalysis. Digital Methods and Literary History* (2013);
Kestemont M. and Herman L., Can Machines Read (Literature)? In: *Umanistica Digitale*, 5. (2019);
Mulligan J., Digital Humanities Application Development in the Cloud. In: *Proceedings of the Humans in the Loop: Enabling and Facilitating Research on Cloud Computing*, pp. 1-5. (2019);
Padilla T., Allen L., Frost H., Potvin S., Russey Roke E., and Varner S., Final Report --- Always Already Computational: Collections as Data (Version 1). (May 22 2019). http://doi.org/10.5281/zenodo.3152935;
Terras M., Baker J., Hetherington J., Beavan D., Zaltz Austwick M., Welsh A., O'Neill H., Finley W., Duke-Williams O. and Farquhar A., Enabling Complex Analysis of Large-Scale Digital Collections. In: *Digital Scholarship in the Humanities*, 33.2, pp. 456–466. (2018)

[74] Graham S., Milligan I., and Weingart S., *Exploring Big Historical Data*. Imperial College Press. (2015). https://doi.org/10.1142/p981

to extrapolate information if needed, and exploring previously unknown sources.

One of the areas where the TM will try to cause a deep impact with innovative research methods is **Simulation Studies (3.1.3)**.[75] Simulation is a scientific technique in which a simplified approximation of a real system (a model) is used to study the dynamics of the system and its evolution over time.[76] Simulation enables testing theoretical models (hypotheses) for cases in

---

[75] Axtell R. L., Epstein J. M., Dean J. S. et al, Population Growth and Collapse in a Multiagent Model of the Kayenta Anasazi in Long House Valley. *PNAS* 99, pp. 7275–79. (2002).
doi:10.1073/pnas.092080799;
Barcelo J., del Castillo F., Simulating Prehistoric and Ancient Worlds. (Springer Berlin/New York) (2016);
Barton, C.M., Complexity, Social Complexity, and Modeling. *Journal of Archaeoligcal Method and Theory*, 21, pp. 306-324*.* (2014).
doi:10.1007/s10816-013-9187-2;
Brughmans T., Hanson J.W., Mandich M.J., Romanowska I., Rubio-Campillo X., Carrignon S., Collins-Elliott S., Crawford K., Daems D., Fulminante F., de Haas T., Kelly P., Moreno Escobar M. del C., Paliou E., Prignano L. and Ritondale M., Formal Modelling Approaches to Complexity Science in Roman Studies: A Manifesto. In: *Theoretical Roman Archaeology Journal*, 2.1, pp. 4-15. (2019).
doi:10.16995/traj.367;
Cegielski W.H., and Rogers J.D., 2016: Rethinking the role of Agent-Based Modeling in archaeology. In: *Journal of Anthropological Archaeology*, 41, pp. 283–298. (2016).
doi:10.1016/j.jaa.2016.01.009;
Epstein J. M., and Axtell R. L., Growing Artificial Societies. Social Science from the Bottom Up. (MIT Press. Boston) (1996).
Ewert U. C., and Sunder M., Modelling Maritime Trade Systems: Agent-Based Simulation and Medieval History. In: *Historical Social Research*, 43, pp. 110-43*.* (2018). doi:10.12759/hsr.43.2018.1.110-143;
Kohler, T. A., Complex Systems and Archaeology. In: *Archaeological Theory Today*, ed: I. Hodder, pp. 93-123. (Polity Press. Cambridge) (2012).
Lake M.W., Trends in Archaeological Simulation. In: *Journal of Archaeological Method and Theory*, 21, pp. 258–287. (2014).
doi: 10.1007/s10816-013-9188-1;
Premo L.S., Agent-based models as behavioral laboratories for evolutionary anthropological research. In: *Arizona Anthropologist*, 17, pp. 91-113*.* (2006).
Romanowska I., Crabtree S. A., Harris K. and Davies B., Agent-Based Modeling for Archaeologists: Part 1 of 3. In: *Advances in Archaeological Practice*, 7.2, pp. 178-84*.* (2019).
doi: 10.1017/aap.2019.6;
Wurzer G., Kowarik K., and Reschreiter H., Agent-based Modeling and Simulation in Archaeology. (Springer Berlin/New York) (2015).

[76] **Human Evolution and Hunter-Gatherers**
Barcelo J. A., Del Castillo Bernal F., del Olmo R., et al, Social Interaction in Hunter-Gatherer Societies. Simulating the Consequences of Cooperation and Social Aggregation. In: *Social Science Computer Review*, 32.3, pp. 417-436. (2014). doi: 10.1177/0894439313511943;
Gravel-Miguel C., and Wren C. D., Agent-based least-cost path analysis and the diffusion of Cantabrian Lower Magdalenian engraved scapulae. In: *Journal of Archaeological* Science, 99, pp. 1-9. (2018). doi: 10.1016/j.jas.2018.08.014;
Mesoudi A., O'Brien M. J., The Cultural Transmission of Great Basin Projectile-Point Technology II: An Agent-Based Computer Simulation. In: *American Antiquity*, 73.4, pp. 627-644*.* (2008);
Premo L. S., and Kuhn S. L., Modeling Effects of Local Extinctions on Culture Change and Diversity in the Paleolithic. In: *PLoS One*, 5.12. *e15582.* (2010);
Romanowska I., Gamble C., Bullock S., and Sturt F., Dispersal and the Movius Line: Testing the Effect of Dispersal on Population Density through Simulation. In: *Quaternary International*, 431, pp. 53-63. (2017). doi: 0.1016/j.quaint.2016.01.016.
Prehistory and Farming Societies
Angourakis A., Rondelli B., Stride S. et al, Land Use Patterns in Central Asia. Step 1: The Musical Chairs Model. In: *Journal of Archaeological Method and Theory*, 21, pp. 405-25. (2014). doi:10.1007/s10816-013-9197-0;
Barton C. M., Ullah I. I., and Bergin S., Land Use, Water and Mediterranean Landscapes: Modelling Long-Term Dynamics of Complex Socio-Ecological Systems. In: *Philosophical Transactions of the Royal Society*, *A* 368 (1931), pp. 5275-97. (2010).

which direct experiments on the real system are not possible due to practical constraints or ethical reasons. Simulation is the only scientific method enabling formal testing of hypotheses regarding socio-cultural and socio-natural systems, such as past societies, against empirical data. However, despite being the main method in science simulation (and formal modelling methods in general), it is severely underused in humanities. The most popular technique is agent-based modelling – a bottom-up approach operating on familiar modelling units such as individual agents and space.[77] This is an incipient field with very promising possibilities, and a great amount of research still needs to be performed. There is, however, a low level of standardisation in terms of model documentation, dissemination, verification and replication. As simulation methods can provide better results with big amounts of reliable data, the TM offers a great opportunity to promote and advance the field. In this sense, it aims to create a user friendly and fully documented library of model building blocks covering all aspects of urban evolution including social, economic and urban planning submodels, and a fully integrated platform for the interface between simulation and data including model development, calibration, validation, documentation, and dissemination by means of a standardized protocol and extensive documentation and training material to encourage reuse and replication.[78]

---

doi: 10.1098/rsta.2010.0193;

Cockburn,D., Crabtree S. A., Kobti Z., et al, Simulating Social and Economic Specialization in Small-Scale Agricultural Societies. In: *Journal of Artificial Societies and Social Simulation* 16.4, pp. 4. (2013);

Griffin A. F., and Stanish C., An Agent-based Model of Prehistoric Settlement Patterns and Political Consolidation in the Lake Titicaca Basin of Peru and Bolivia. In: *Structure and Dynamics,* 2.2. (2007).

Classical Studies

Brughmans T., and Poblome J., Roman Bazaar or Market Economy? Explaining Tableware Distributions in the Roman East through Computational Modelling. In: *Antiquity*, 90.350, pp. 393-408. (2016). doi:10.15184/aqy.2016.35;

Graham S., Networks, Agent-Based Models and the Antonine Itineraries: Implications for Roman Archaeology. In: *Journal of Mediterranean Archaeology*, 19.1, pp. 45-64. (2006). doi:10.1558/jmea.2006.19.1.45.

Settlement and Urbanism

Crema E. R., A Simulation Model of Fission – Fusion Dynamics and Long-Term Settlement Change. In: *Journal of Archaeological Method and Theory*, 21, pp. 385-404. (2014). doi:10.1007/s10816-013-9185-4;

Snyder J.R., Dilaver O., Stephenson L.C., Mackie J. E., Smith S.D., Agent-Based Modelling and Construction – Reconstructing Antiquity's Largest Infrastructure Project. In: *Construction Management and Economics*, 36.6, pp. 313-327. (2017). doi:10.1080/01446193.2017.1403639.

Protohistory and History

Ewert U. C., Sunder M., Modelling Maritime Trade Systems: Agent-Based Simulation and Medieval History. In: *Historical Social Research*, 43, pp. 110-43. (2018). doi:10.12759/hsr.43.2018.1.110-143;

Kowarik K., Reschreiter H., and Wurzer G., Modeling a mine. Agentbased Modeling, System dynamics and Experimental Archaeology applied to the Bronze Age Saltmines of Hallstatt. In: *Mining in European History* (SFB HiMAT Innsbruck), pp. 44-53. (2010);

Rubio Campillo X., Cela J.-M., and Hernàndez Cardona F. X., Simulating Archaeologists? Using Agent-Based Modelling to Improve Battlefield Excavations. In: *Journal of Archaeological Science*, 39.2, pp. 347-56. (2012). doi:10.1016/j.jas.2011.09.020.

[77] Heppenstall A.J., Crooks A.T., See L.M., Batty M., Agent-based models of geographical systems. Springer Science & Business Media (2011);

Raimbault J., Cottineau C., Texier M.L., Néchet F.L. and Reuillon R., 2018. Space matters: extending sensitivity analysis to initial spatial conditions in geosimulation models. (2018). arXiv preprint arXiv:1812.06008.

[78] Batten D.F., *Discovering artificial economics: How agents learn and economies evolve*. Routledge. (2019);

Groeneveld J., Müller B., Buchmann C.M., Dressler G., Guo C., Hase N., Hoffmann F., John F., Klassert C., Lauf T., and Liebelt V., Theoretical foundations of human decision-making in agent-based land use models–A review. In: *Environmental modelling & software*, 87, pp.39-48. (2017);

Huang Q., Parker D.C., Filatova T. and Sun S., A review of urban residential choice models using agent-based modeling. In: *Environment and Planning B: Planning and Design*, 41.4, pp.661-689. (2014);

Digital Methods (3.1.4) aims at promoting a critical approach to the use of digital research infrastructures, tools and data. It means an "update" of classical hermeneutics in the field of humanities to the digital age[79] towards *digital hermeneutics*, reflecting the whole life-cycle of a humanities research process: from the developing of a research question, to the retrieval of information, source criticism, analysis and interpretation, to developing an argument and producing a narration It thus comprises a set of skills: infrastructure criticism, algorithmic criticism,[80] data criticism,[81] tool criticism,[82] and interface criticism.[83] It also interrogates the innovations allowed by "distant reading":[84] 1- these innovations can be used heuristically, whereby the patterns observed lead to new hypotheses on the phenomenon under investigation, which are subsequently analysed with traditional, interpretative methods. 2- The collection of big data can be used to empirically test existing assumptions based on smaller sample data. 3- They allow for the combination of different types of data and thus for more complex analyses.

3.2. DISCIPLINES focus on how the TM interacts with specific academic disciplines and traditions. **History (3.2.1)** will be most influenced by the TM, as the data uncovered will enable the application of research methods that were only partly feasible,[85] with a focus on large amounts of quantitative information. Although computational methods in historical research go back to the '70s and '80s, they are still underdeveloped and mostly marginal.[86] Traditional geographical (mainly national) and chronological boundaries still define most historical inquiries, although there have been several attempts to undermine those divisions. The TM will change the field by providing access to an enormous amount of previously hidden information in the form of linked data, highly increasing discoverability of new facts. The TM will also create the condition for a renewed impetus in "*longue durée*" analysis, i.e. studies that consider the evolution across a long-term frame. The innovations enabled by the access to Big Historical Data are to 1. Pose new questions about the past; 2. Empirically test existent assumptions based on much smaller

---

Pumain D., (ed.), *Geographical Modelling: Cities and Territories*. (2020);
Pumain D., (ed.), *Theories and Models of Urbanisation*. In: *Lecture Notes in Morphogenesis*. Springer (2020);
Sanders L., Pumain D., Mathian H., Guérin-Pace F., and Bura S., SIMPOP: a multiagent system for the study of urbanism. In: *Environment and Planning B: Planning and design*, 24.2, pp. 287-305. (1997);
Youngman P.A., Hadzikadic M., Complexity and the human experience: modeling complexity in the humanities and social sciences. CRC Press. (2014).

[79] Zundert J. J. van, Screwmeneutics and Hermenumericals: The Computationality of Hermeneutics. In: *A New Companion to Digital Humanities.* Ed. Schreibman S., Siemens R. G., and Unsworth J. (2016);
Verhoef J., and Wevers M., The Digital Humanities Cycle: Hermeneutics, Heuristics, and Source Criticism in a Digital Age. In: *The Digital Humanities cycle: hermeneutics, heuristics, and source criticism in a digital age* (2015);
Armaselu F., and Heuvel C. v. d., *Metaphors in Digital Hermeneutics: Zooming Through Literary, Didactic and Historical Representations of Imaginary and Existing Cities*. 11.3 (2017);
Romele A., Severo M., and Furia P., Digital Hermeneutics: From Interpreting with Machines to Interpretational Machines. In: *AI & Society* (2018). doi:10.1007/s00146-018-0856-2

[80] Ramsay S., *Reading Machines: Toward An Algorithmic Criticism.* Urbana: University of Illinois Press (2011).

[81] Strasser B. J, and Edwards P. N., Big Data Is the Answer … but What Is the Question. In: *Osiris* 32.1 (2017). doi:10.1086/694223

[82] van Es K., Wieringa M., and Schäfer M. T., Tool Criticism: From Digital Methods to Digital Methodology, In: *The 2nd International Conference.* ACM Press (2018).

[83] Warwick, C., Interfaces, Ephemera, and Identity: A Study of the Historical Presentation of Digital Humanities Resources. In: *Digital Scholarship in the Humanities Digital Scholarship Humanities* (2019). doi:10.1093/llc/fqz081

[84] Underwood T., A Genealogy of Distant Reading. In: *Digital humanities quarterly* 011.2 (2017).

[85] Demantowsky, M., What is Public History: International Perspectives. (2018). DOI10.1515/9783110466133-001.

[86] Jockers M. L., Macroanalysis: Digital Methods and Literary History. University of Illinois Press (2013).; Mounier P., Les humanités numériques: une histoire critique. Paris: Maison des sciences de l'homme (2018).

datasets; 3. Do more complex analyses, because a) we have larger, more comprehensive datasets (as TM provides a solution for the current fragmentation of sources) and b) we can combine different types of data in one analysis, allowing us to explore explanations that are difficult to analyse by hand at present (such as statistical relations between socio-demographic factors and other data on e.g. migration, language use, cultural consumption, or voting behaviour, etc.).

In **Language and Literature (3.2.2)** the advances in NLP (Task 2.2) will allow an improved approach to old language and literature. The studies on language evolution can incorporate new statistical techniques working with bigger amounts of data.[87] Distant reading in literature, that is usually only available for the 19th century,[88] will be able to be carried out for previous European periods. These methods will be combined and complemented by traditional philology and literary history. The TM can aim for new holistic perspectives on literary texts, which will be contextualized and linked into the global semantic network and knowledge base of the TM: comprehensive empirical data on processes of language change, including the spatio-temporal spread of changes and the role of individual speakers/authors, validation and revision of existing theoretical explanations, development of new theories, and the establishment of novel paradigms for corpus-based research which enable systematic quantitative-qualitative analysis especially of specialized and historical corpora, achieving statistical significance with far smaller data sets than purely quantitative techniques.

In **Archaeology (3.2.3),** the TM aims to provide significant contributions to the domain. In the wake of the "Malta" European legislation (on the protection of archaeological heritage),[89] a massive expansion of digital archaeological data took place in the past decade,[90] mostly in the framework of "commercial archaeology", producing an enormous amount of so-called grey literature. Currently, national data-infrastructures are prevalent, with different terminologies, data standards, and access rules.[91] Archaeology increasingly uses non-invasive technologies (from Digital Height Models to geophysical reconnaissance techniques). The TM will benefit the field in two ways: on the one hand, by developing new non-invasive acquisition techniques, and on the other hand by providing access to large scale data from at least the urban areas of Europe, as well as innovative 3D models and visualisations.

In the fields of **Art History and Media Studies (3.2.4),** Many databases have been developed in the past, but with very heterogenous models and mostly proprietary software, yet many institutions have digitised their holdings, focusing on the most famous objects. Some projects, such as Europeana, have improved the situation by providing a framework for European art digitisation projects. Standards like CIDOC CRM and LIDO are used to deal with the

[87] Bode K., Reading by Numbers: Recalibrating the Literary Field. New York: Anthem Press (2012). http://www.anthempress.com/reading-by-numbers;
Pang B., and Lee L., Opinion Mining and Sentiment Analysis. (2008). http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[88] Moretti F., Distant reading. Konstanz (2016).

[89] https://www.coe.int/en/web/conventions/full-list/-/conventions/rms/090000168007bd25

[90] http:/vwhl.squarespace.com; https:/www.beazley.ox.ac.uk/index.htm; http://numismatics.org/resources/

[91] https:/caa-international.org/proceedings/published/ and Conolly J., Geographical Information Systems in Archaeology. 3rd printing. Cambridge: Cambridge University Press (2008). http://dx.doi.org/10.1017/CBO9780511807459;
Collar A., Coward F., Brughmans T., and Mills B. J., Networks in Archaeology: Phenomena, Abstraction, Representation. In: *Journal of Archaeological Method and Theory*, 22.1, pp. 1-32. (30. January 2015). https://doi.org/10.1007/s10816-014-9235-6

complexity of the data. In art history and digital humanities the field of digital art history is establishing.[92] The CLARIAH-NL research infrastructure has developed a first solution for accessing and analysing in-copyright or privacy-sensitive audio-visual collections, including automated speech recognition for quantitative content analysis (https://mediasuite.clariah.nl/). The TM will contribute to the field in an array of ways: firstly, it will make guidelines available to GLAM institutions for digitisation (quality, techniques, etc.) and storage (cf. Task 1.3) that fit into the existing institutional landscape and aim at complementing it. Secondly, it will develop technologies to improve digitisation and analysis of images, audio and video (Task 1.1). Lastly, it will also create searchable databases with many objects and metadata, connected to previous infrastructures.

In **Geography and Demography (3.2.5)**,[93] quantitative approaches are established in the field and large amounts of digital geodata at different geographic scales are continuously produced by European countries using different data-standards and different policies with regard to access. The TM will offer integrated and flexible access to historical geodata at European level and allow highly efficient and fine-grained spatial analyses of key-demographic indicators (e.g. social and professional topographies, migration, mortality and health etc.).

In **Musicology (3.2.6)** the TM will improve the digitisation and computer analysis of music scores which will enable distant reading of the materials and a better understanding of musical evolution in Europe. The 3D digitisation of musical instruments and 2D digitisation of paintings and drawings containing musical instruments will also provide useful tools for research.

Digital Humanities (3.2.7),[94] a growing subject in Europe[95] and the TM is an essential part of that growth, as an exemplary case of joining developments in computer science and CH.

---

[92] Kuroczyński P., Bell P., and Dieckmann L., Computing Art Reader: Einführung in die digitale Kunstgeschichte. In: *Computing in Art and Architecture* (2018). https://doi.org/10.11588/ARTHISTORICUM.413;
Baca M., Helmreich A., and Gill M., Digital Art History. In: *Visual Resources*, 35:1-2, pp. 1-5. (2019). DOI: 10.1080/01973762.2019.1556887;
Kohle H., *Digitale Bildwissenschaft*. Heidelberg University Library (2013). https://doi.org/10.11588/artdok.00002185

[93] Glasze G., Geoinformation, Cartographic (Re) Presentation and the Nation State: A Co-Constitutive Relation and Its Transformation in the Digital Age. In: Kohl, U. (Ed.): *The Net and the Nation Sate. Multidisciplinary Perspectives on the Internet Governance*. Cambridge: Cambridge University Press, pp. 218-240. (2017);
Westerholt R., Mocnik F.-B. & Zipf A. (Eds.): On The Way To Platial Analysis: Can Geosocial Media Provide The Necessary Impetus? In: *Proceedings Of The First Workshop On Platial Analysis (Platial '18)*. Heidelberg: Zenodo (2018);
Paterson L. L., & Gregory I. N., Representations of poverty and place: using geographical text analysis to understand discourse. (2019);

[94] Kirschenbaum M., What Is “Digital Humanities”, And Why Are They Saying Such Terrible Things About It? In: *differences* 25, pp. 46-63. (2014). doi:10.1215/10407391-2419997

[95] Berry D. M., Post-digital humanities: computation and cultural critique in the arts and humanities. In: *Educause* 49, pp. 22-26. (2014). http://www.educause.edu;
Berry D. M., and Fagerjord A., *Digital humanities: knowledge and critique in a digital age*. Cambridge Malden, MA: Polity (2017);
Fiormonte D., Towards a Cultural Critique of the Digital Humanities. In: *Historical Social Research / Historische Sozialforschung*, 37, pp. 59-76. (2012). https://www.jstor.org/stable/41636597;
Fish S., The Digital Humanities and the Transcending of Mortality. In: *The New York Times* “Opinionator”. (2012). https://opinionator.blogs.nytimes.com/2012/01/09/the-digital-humanities-and-the-transcending-of-mortality/;
Barget M., “The real problem with Digital Humanities” – critical approaches for further discussion. Revolts as Communication. (2018). https://revolt.hypotheses.org/1848;
Champion E. M., Digital humanities is text heavy, visualization light, and simulation poor. In: *Digital Scholarship in the Humanities* (2016). fqw053. https://doi.org/10.1093/llc/fqw053

Connecting with ongoing DH projects and institutions will be a priority for the TM.

Finally, **Urban Studies (3.2.8)** will benefit greatly from the TM. At the moment the "Smart Cities Initiative" (SCI), innovative approaches in urban policy, planning & development, make traditional networks and services more efficient for urban users (citizens, policy makers, businesses, etc.) through the use of digital and telecommunication services. The SCI generate massive amounts of data, but have limitations, mainly focusing on the present, due to data availability. TM will provide the longitudinal perspective, that represents the "collective, long-term memory" of these cities – from historical depth to present-day data collections. TM will enhance data collecting, processing and integration, by enriching with unconsidered historical data collections, for instance. SCI are usually set-up by private companies, which creates problems relating to algorithm use, privacy and data ownership. TM will set-up a FAIR open data - open access data infrastructure, according to the principles of data being Findable, Accessible, Interoperable and Reusable. TM will collaborate with municipalities and public-private initiatives to make urban data accessible and connected to the Big Data of the Past while assisting in the development of plans and strategies.

Targeted Achievements

After assessing the state of the art in the various Science and Technology subdomains that are relevant to the Time Machine, it becomes clear that a number of specific innovations and improvements must be targeted to realize the Pillar's overall objective. The urgency and difficulty of these challenges varies somewhat across different domains. Nevertheless, while an appropriate prioritisation strategy must be adopted in the roadmap, all of these targeted achievements should be considered crucial milestones in the development of the TM.

Each domain defined in the scientific taxonomy for Pillar 1 (Data, Computing, Social Sciences and Humanities) has a series of Targeted Achievements (A). In order to accomplish them, some milestones (M) have been established, many of them in the form of RFC. The targeted achievements and milestones are listed and explained below. Each of these Targeted Achievements involves work of particular areas of expertise according to the taxonomy for Science and Technology developed above, which are also mentioned in approximate order of relevance for the specific goal.

1. DATA

A1.1. Digitisation Hubs

M1.1.1– **RFC for Digitisation Hubs**: In order for the Digitisation Hubs to be implemented, standards in terms of resolution, file formats, and metadata during acquisition need to be defined. These must be consensual and simple, in order to be easily implemented and to fit into existing practices. The RFC also needs to evaluate relevant technologies and recommend affordable technology that does not damage the objects, providing the best possible results at the same time. We aim to distribute cheap technology on a large scale using e.g. open design hardware. More costly and dedicated scan methods such as scan robots and tomographic methods should be available in dedicated specialized centres spread across the European Union such that their services are available to a maximum number of users. The objective of Pillar 3 of achieving cheap and wide-spread digitisation should be a priority in this RFC.

M1.1.2– **Implementation of Digitisation Hubs**: The Digitisation Hubs designed according to the results of M1.1.1 will start functioning. In the first stage, we aim predominantly at a wide-spread use of standardised and inexpensive technology. A review process and user consultation should take place 3 months after the beginning of the first stage, and then periodically, the aim of the review being to identify weaknesses in scanning recommendations, hardware, and software post-processing.

M1.1.3– **RFC on New Scanning Technologies.** Cutting-edge technologies, such as automatic scanning machines with low human supervision, scanning robots and solutions for scanning films and books without the need to unroll/open them, should be considered and fostered by the TM. A specific scheme to incentivize these technologies will be created. We aim at an appropriate mix of dedicated specialized scanning centres and at the development of mobile special use hardware, e.g. mobile CT scanners that are mounted on trucks.

Main Areas Involved: 1.1

A1.2. TM Box (Servers)

M1.2.1– **RFC for TM Box**: The features of the distributed storage system where the Data Graph is to be hosted will be discussed by the community for this milestone. Important issues are the technical server infrastructure, the compliance with international standards, the creation of a system to prove trustworthiness via certification processes, de-duplication methods leveraging pattern-recognition across large datasets, and the implementation of digital observatory and digital archive layers. Furthermore, connection to long-term storage, e.g. DNA storage, and the selection of the most important data to be stored in such archives, is a pertinent design question.

M1.2.2– **Implementation of TM Box**: The data stored in different individual storage systems up to this point (LTM and partner institutions) are copied or linked to the TM Box and the correct access is assessed.

Main Areas Involved: 1.3

A1.3. TM Data Graph

M1.3.1– **RFC on Technical Charter:** The goal of the Charter is to guarantee a first level of standardisation for data and processes, in order to remain light and useable by the most, the charter also encourages the use of universal and open interfaces and references that do not need central coordination. Definition of guidelines and standards to follow need to be set up regarding formats and protocols to store and query data. Proper data management and curation enable interconnection and involvement of diverse research disciplines and therefore provide an excellent environment for boosting innovation. Data will be made available with trustworthiness and FAIR (Findable – Accessible – Interoperable – Reusable) principles in mind. They will include content (primary data), metadata and derivates (secondary data) as well as externally linked data. Primary Data should be preserved in a Digital Archive with persistent identifiers, usually called a Trusted Digital Repository. Secondary Data should be stored in the research infrastructure with data versioning and full provenance information. Linked Data should be available in Linked Open Data Cloud (LOD). All data should be stored in these FAIR repositories with possibility to get data in and out in the standardised way.

Main Areas Involved: 1.2

M1.3.2– **RFC on Digitisation Priorities and Data Selection:** A typology of digitisation interventions will be established, separating a. collections that can be moved and processed in digital hubs (large, non-fragile collections), collections or objects that need local intervention (e.g. very fragile document, statues, buildings), b. process that can be performed by volunteers using mobile technology (e.g. scanning campaign across cities, on-the-fly digitisation in reading rooms), processes that can be performed using robots and drones, etc.

2. COMPUTING

A2.1. Interface for Annotation

M2.1.1– **User Studies of Current Annotation Platforms**: Good quality annotation is key to create a linked Data Graph for the TM. In order to produce human annotations of quality, a proper interface is required. As many such interfaces currently exist, an assessment of the landscape is a prerequisite for the creation of a new TM interface.

M2.1.2– **RFC on Annotation**: Definition of the annotation protocols used for the document of TM Data Graph. This RFC will be used by exploitation platforms that must allow for easy but complex annotation, complying with the standards set for data modelling (M.1.3.1). The principles of human-computer interaction (taxonomy task 2.4) and previous user studies (M.2.1.1) will inform the development of the annotation tools.

Main Areas Involved: 2.4.1 / 2.4.7 / 1.2.4

A2.2. User Interface

M2.2.1– **User Studies of Current Platforms for Historical Data**. Users of the TM will be able to access the data and materials produced by the TM through user interfaces. Some of them will be developed by the LTMs for their own purposes, but a central interface, as well as templates for the LTMs that require them, must be elaborated. The first step towards this goal is an assessment of the current interfaces being used in the LTMs and other projects concerning the digitisation of CH.

Main Areas Involved: 2.4

A2.3. Natural Language Processing Tools for Older Language Variants

M2.3.1– **RFC for Classification and Planning of Languages to Address**. The TM will handle documents in multiple European languages and dialects. Some of them might be more complicated to address than others due to pre-existing tools for modern variants or availability of materials. A working plan of NLP tools development should be conceived by taking the materials, the locations of the LTMs and the Digitisation Hubs, and the features of the languages into consideration.

M2.3.2– **RFC for Named Entity Recognition**. Based on the plan outlined in M2.5.1, the community will develop tools for named entity recognition in older European languages and variants. The results of the tagging of entities will feed the Dark Data Graph with new information.

M2.3.3– **RFC for Orthographic Normalisation**. Based on the plan outlined in M2.5.1, the community will develop tools for orthographic normalisation of older European language variants. The results will improve the search functionality of the databases and be useful for M2.5.4.

M2.3.4– **RFC for Machine Translation**. Existing algorithms for machine translation will be adapted to older language variants of European languages as outlined in M2.5., taking advantage of the results of M2.5.2 and M2.5.3

Main Areas Involved: 2.2 / 2.3

A2.4. Digital Content Processor

M2.4.1. RFC for Digital Content Processor Level 1: Using Machine Learning from existing annotated data, the Digital Content Processor level 1 will be able to label mentions of entities. Results from M2.3.2 will be essential for this development.

M2.4.2. RFC for Digital Content Processor Level 2: Level two will be able to create labels to establish relationships between entities to create linked data that improves the Data Graph.

M2.4.3. RFC for Digital Content Processor Level 3: Level three is able to create re-useable models, that generalize from few observations and contribute to possible understanding of the patterns behind the available data.

Main Areas Involved: 2.3.2 / 2.3.6 / 2.38 / 2.3.9 / 2.2.3

A2.5. TM Engines

M2.5.1– RFC for TM APIs. Algorithms and software integrated into the Time Machine need to be able to communicate with each other. Thus, a definition of joint APIs is required. It is likely that TM Services are built on top of REST interfaces. In order to match TM's needs these will have to be adapted towards the need of large-scale machine learning. A probable addition is the option to provide gradient information of a specific module that is integrated using the API, for example. This way remote services can also be integrated into large-scale training processes.

M2.5.1– RFC for Large-Scale Inference Engine. The Large-Scale Inference Engine will support fact-based reasoning and logical deduction. It needs to provide a dedicated API to generate new insights from data. Therefore, it needs to enable addition of new evidence, hypothesis checking, and retrieval of related data nodes.

M2.5.2– RFC for 4D Simulator. The 4D Simulator needs to interface with the Time Machine Data Graph so that virtual worlds can be derived from known evidence. At the first stage, this will imply the loading of pre-configured world models and objects. At later stages, the generation and adaptation of existing models according to new evidence will also be the focus of the 4D Simulator. Furthermore, another focus will be the generation of virtual agents inhabiting the generated world of the 4D Simulator. An appropriate modular API has to be designed with respect to these requirements.

M2.5.3– RFC for Universal Representation Engine. The Universal Representation Engine will make use of the earlier discussed Universal Representation Space. The aim of its use is to support creative and connotative research methodologies. Therefore, APIs have to be defined that enable conversation of e.g. images to text, description to 3D objects, or even maps to virtual cities and vice versa. Again, appropriate design choices have to be made to design APIs and to enable intercommunication.

M2.5.4– RFC for Mirror World Prototyping. Definition of the implementation strategy for first working protoype of Mirror World using the TM engines. This first prototype is likely to be developed on the most advanced LTM. This RFC will define how to safely experiment with this technology.

M2.5.5– RFC for Mirror Technical Standards. This RFC defines the specific technical standards needed for the Mirror World extension, based on the experience developed with the Mirror World prototype

Main Areas Involved: 2.6 / 2.5 / 2.3

A2.6. Automatic Text Recognition

M2.6.1– Call on Text Recognition: A call to researchers and TM partners working in Text Recognition will be issued to use the digitised documents of the TM to improve the existing models.

M2.6.2– RFC on Text Recognition and Processing Pipeline: Following the results of M2.4.1, general models for text recognition should be created, i.e. models that work for the largest number of similar documents possible, so that no new models need to be trained to process texts in almost any European script.

Main Areas Involved: 2.1.1 / 2.3

A2.7. Automatic Graphic Document Recognition

M2.7.1.– RFC for Map Recognition : Using the digitised materials of the TM, the methods and results of automatic map recognition should be improved. Depending on the results of the first RFC on map recognition, another case should probably follow.

M2.7.2.– RFC for Music Scores Recognition : Using the digitised materials of the TM the methods and results of automatic music scores recognition should be improved.

Main Areas Involved: 2.1.2 / 2.3

3. SOCIAL SCIENCES AND HUMANITIES

A3.1. Improved Integration of Quantitative Methods in Historical Research

M3.1.1.– Call for Quantitative Historical Research with TM Data Graph (1-3). The TM should incentivize and create a framework for researchers in historical subjects (history, literature, art, musicology, etc.) to use the TM Data Graph to perform quantitative historical studies as well as facilitating a *longue durée* perspective. This will be achieved by organising dedicated conferences and open calls for papers. There should be at least three organised phases in the strategy to enhance historical research with the Big Data of the Past. The first phase will happen before the development of the TM tools for historical research (M3.1.1), in order to better assess the needs of the scholarly community. The second will take place right after the publication of those tools, so that researchers can test and use them. The third call will materialize in the final phase of the project.

M3.1.2.– RFC TM Tools for Historical Research. To engage researchers in social sciences and humanities is to productively use the Big Data of the Past, and the TM can offer researchers a series of tools that facilitate analyses. These tools will be enhanced by the Digital Content Processor and the Simulation Engines, which will enable scholars to work with historical data in an unprecedented way.

Main Areas Involved: 3.1.1 / 3.1.2 / 3.1.4 / 3.2.1 – 3.3.4

A3.2. Successful Historical Simulations using the TM Data graph

M3.2.1.– Call for Agent-based Simulation Using Linked Data. Agent-based simulation can be achieved, in a first stage, using the information from the Data Graph and from models outside the TM. A call for teams working on simulation studies that want to test models with information from the TM should be issued to assess the quality of the data and the existing models.

M3.2.2.– RFC for Improved Simulation Using TM Simulation Engines. Researchers will

be able to use the TM simulation engines (A2.5) to perform simulations studies, without having to rely on external models and tools. The simulation engines have the capacity to improve the performance and the reach of computational simulations for historical research.
Main Areas Involved: 3.1.3 / 3.2.1

Milestones

An approximated timeline for the accomplishment of the milestones in the course of the next 10 years has been developed considering the requirements of each of them and can be found below. The most immediate milestones are those concerning “Data” as these milestones form the basis to digitise, share, and store the Data Graph which is a source material needed by the other areas to accomplish their targeted achievements. The dates are approximate seeing as the barriers, difficulties and success rates can only be partially assessed for a 10-year period and will depend on decisions made by the TM operation structure.

Thematic Area / Milestones	Time										
	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
Data											
1. Digitisation Hubs											
RFC Digitisation Hubs											
Implementation Digitisation Hubs											
RFC New Scanning Technologies											
2. TM Box											
RFC TM Box											
Implementation TM Box											
3 TM Data Graph											
RFC on Technical Charter											
RFC on Digitization Priorities and Data Selection											
Computing and Artificial Intelligence											
1. Annotation Interface											
User Studies of Current Annotation Platforms											
RFC on Annotation											
Annotation Interface Implementation											
2. User Interface											
User Studies of Current Platforms for Historical Data											
User Interface Implementation											
3. Natural Language Processing (NLP) Tools for Older Language Variants											
RFC on Classification and Planning for Languages to Address											
RFC for Named Entity Recognition											
RFC for Orthographic Normalisation											
RFC for Machine Translation											
4. Digital Content Processor (DCP)											
RFC for DCP Level 1											
RFC for DCP Level 2											
RFC for DCP Level 3											
5. Simulation Engines											
RFC for TM APIs											
RFC for Large-Scale Inference Engine											
RFC for 4D Simulator											
RFC for Universal Representation Engine											
RFC for Mirror World Prototyping											
RFC for Mirror World Technical Standards											
6. Automatic Text Recognition											

Call on Text Recognition											
RFC on Text Recognition and Processing Pipeline											
7. Automatic Graphic Document Recognition											
RFC for Map Recognition											
RFC for Music Scores Recognition											
Social Sciences and Humanities											
Improved Acceptance of Quantitative Methods in Historical Research											
Call for Quantitative Historical Research with TM Data Graph (1)											
RFC TM Tools for History Research											
Call for Quantitative Historical Research with TM Data Graph (2)											
Call for Quantitative Historical Research with TM Data Graph (3)											
Successful Historical Simulations using TM Data Graph											
Call for Agent-based Simulation Using Linked Data											
RFC for Improved Simulation Using TM Simulation Engines											
Implementation											
Yearly Open Calls											

Proposed Methodologies

The devising of both a state of the art and a vision for each field in our taxonomy enables us to fully develop the TM Science and Technology Road Map. As demonstrated in the milestone table, there is one main instrument to achieve the proposed objectives, namely the Requests for Comments (RFC). The structure and details of the RFC will be explained in the roadmap for Pillar 2 "Operation". RFCs revolve around a publication format based on openness and accessibility, which enhances collaboration and reduces operational costs. A strong and committed community is able to achieve great outcomes using this format, as can be attested from the development of the internet. For this reason, many of the milestones proposed in this Pillar have the form of RFC with particular objectives. The RFCs are accompanied by a set of fundamental research questions that need to be clarified by scientific project work (e.g. user studies required to prepare the development of a RFC). These projects need to be funded directly by the Time Machine Project. In this category, calls for papers and conferences will enhance communication, help clarify the goals, and assess the progress in different areas.

Once individual RFCs are developed, we propose to implement the roadmap via a modular design during the coming 10 years, in which various calls for proposals aim to attract bottom-up research proposals targeting specific milestones within a pre-specified time frame. These research initiatives can range from macro- to micro-level funding initiatives, and from both local and European funding initiatives. In order to increase the impact of TM and to create synergy with the many European efforts that are already in place, we propose to connect funding from Time Machine to base projects that have already been acquired by the applicants to develop their project idea either from national, European, or industrial resources. TM will then provide "**on top funding**" to enable the correct implementation of APIs and methodologies proposed in the RFCs. Additional synergy is created as TM funding can be linked to preferred data and license models enforcing open access policies.

The TMO will also create a network that enables collaboration of individuals and institutions working towards particular goals. Partner Institutions will develop technologies in their expertise, using the Time Machine as a hub that can foster their collaboration. The Local Time Machines will use their own institutional infrastructures and take advantage of the general Time Machine architecture when required. The periodical meetings (such as the annual TM Conference) will be essential to share and evaluate the progress and the achievement of the proposed milestones. The Time Machine will receive input and suggestions from the Local Time Machine experiences, the users, and the academic community, especially in humanities and social sciences, in order to understand the necessities and demands of the public.

Based on discussions that were held during the Pillar 1 Workshop in Dresden, a shared interest emerged towards a research plan that would translate itself into "**challenges**" or "**shared tasks**", which are a popular scientific format in the more exact sciences, in particular in computer science. These challenges or tasks would allow us to formalize and concretize the scientific challenges that would be crucially required by the anticipated functionality of the so-called TM Box (a Pillar 2 concept). The principle of annual evaluation campaigns in the light of these challenges could serve as a useful array of key performance indicators, to track the project's progress towards our shared, abstract goals, but also to identify new issues along the road, which cannot be foreseen at this stage.

Challenges will therefore play a crucial role in the “agile unfolding” of the scientific component of the TM. We will devise a funding scheme around a priority-ranked, structured list of (example) challenges, whereby the prototypes of submitted software (during the annual campaigns) can be handed off to a TM deployment team once the desired performance threshold of a piece of technology is satisfactory. The results of the yearly evaluation campaigns can always be fed into the RFCs, ensuring dynamicity on the base of annual updates and reviews. This procedure could evolve into a vital management instrument for the TMO, by providing an active governance scheme lead by dedicated contact persons who are experts in their field.

Key Performance Indicators

The Key Performance Indicators (KPI) are organised according to the taxonomy, although with different levels of granularity. In some areas, each topic is specific enough to have their own KPI, while other areas have only one set of KPIs. At the same time, some topics can be easily measured in many distinctly quantified ways, while others can only be quantified as a vague indicator of the actual state of affairs.

Data Acquisition: Number, diversity, and types of objects digitised and quality of the digitisation.

Data Storage: Data-loss probability; overall system operating cost. Number of total / well-formed & valid file formats within an archive; ratio file formats in archive / available tools for analysis.

Data Modelling: Publication of the TM guidelines for data and metadata as part of the TM Official Components. Scope of the integration with other initiatives. Number of certified digital archives. Speed of development, speed of adoption, percentage of assets making use of models.

Text Recognition: Accuracy in terms of Word/Character Error Rate (CER/WER). Variety in languages and type of documents. Free available tools.

Graphic Document Processing: Accuracy in terms of false positive rate (FPR). Variety type of documents. Free available tools.

Indexing and Retrieval: Number of public and private institutions making their collections searchable. Number of searches carried out by final users on these collections. Number of validated interconnected documents via search engines. Classical performance indicators (precision, recall, mean average precision, etc.) on cross domain and multimodal collections. Performance indicators versus required memory and search time. User studies.

Understanding and Interpretation: Accuracy and AUC for classification. Recall@{1,5,10} for metric learning and localisation. Distance in meters for localisation. User studies.

Recognition and Detection: For classification, accuracy and AUC; for detection, average precision. Intersection over Union (IoU).

Person & Face Identification: Face detection performance in different content domains (as precision/recall, MAP) compared to human (in identification and verification tasks). Face recognition performance in different content domains, across persons’ lifetimes (as precision/recall, MAP).

Audio Recognition and Transcription: WER for speech recognition, Number of institutions and media providers that make their archives searchable. Number of searches carried out by the final users of the archives. Number of enriched archives.

Machine Learning and AI: Speed and efficiency of technologies. Performance on large-scale benchmarks. CH bots accuracy in human understanding, language generation and human understanding. Avoidance of biases. User studies.

Computer Graphics: Faithful Renderings of historic artefacts in their original context, in real-time and thus applicable for VR and AR. Quality of visualisation, supported platforms. Tracking offset. Perceived lighting artefacts. User studies.

Natural Language Processing: Error rate of methods (accuracy, F1 score, BLUE scores, etc.). Language and variants where they are effective. User studies.

Human-Computer Interaction and Visualisation: Results of user studies. Number of users of the TM interfaces.

Humanities and Social Sciences: Engagement of academia and research (Bibliometrics, Alt Metrics) with the TM through mentions in journals and books, initiatives and projects using the TM data or infrastructure.

FUNDING SOURCES

Many of the technologies are already being developed using a variety of funding sources in the involved institutions from European, national, and industrial resources (a list of European and national founding sources is omitted at this point).

As yet, none of these funding sources are able to support a large-scale project such as the Time Machine. Furthermore, as project-driven research is typically limited in budget, there is no incentive for projects to implement TM APIs. However, a central, large-scale funding mechanism is required to implement the Time Machine as a whole.

Pillar 1 is particularly relevant to research themes considered in the different parts of Horizon Europe, as shown in the table below.

Time Machine Pillar 1	Horizon Europe		
	Pillar I	**Pilar II**	**Pillar III**
Data	x	x	
Computing	x	x	
SSH	x	x	x

Most activities of Pillar 1 relate to the Horizon Europe (HE) clusters Digital, Industry and Space and Culture, Creativity and Inclusive Society (HE Pillar II). For the former cluster, Pillar 1 will develop multimodal historical and geographic datasets of an unprecedented semantic complexity that will give a new impetus for big data research, methods and application fields. Due to its transversal nature as a backbone for other critical technologies, Time Machine is expected to contribute to many other areas of HE, including AI, big data and machine learning in Pillars I and II of HE, and give a strong boost to SSH topics across Pillar III of HE.

By making use of the proposed "on top funding" methodology referred to in the Proposed Methodologies section, research labs, universities, and private companies are incentivized to integrate in the Time Machine Project. This will enable a combining of developments already in progress by their existing funding scheme within the grand vision of the Time Machine. Yet, to fully develop the required technologies for the Big Data of the Past, a series of specific funding for the development of RFCs, user studies, and light-house projects is needed. This can only be achieved using a large-scale research initiative.

STAKEHOLDERS TO BE INVOLVED

The Time Machine concept has brought together a very broad partnership of leading European academic and research organisations, cultural heritage institutions and private enterprises. The members of this unique alliance are fully aware of the huge potential of digitisation and the very promising new paths for science, technology and innovation that can be opened through the information system that we propose to develop, based on the big data of the past.

For Pillar 1, the stakeholders to be involved are:

- Members of pan-European scientific associations like European Open Science Cloud and SSHOC
- Large-scale research initiatives in the area of AI, HPC and robotics
- Professional organisations for historians / archivists / libraries / museums
- International umbrella institutions providing the necessary authoritativeness for the TM approach, such as ICOM for museums, IFLA for libraries, ICA for archives
- Owners of legacy material and objects

The Consortium has approached and received feedback and intensions to support the proposed work programme by an extensive part of these stakeholders.

Below is a list of new members that besides the Consortium Partners have been involved in the Pillar 1 roadmap:

- 3Dkosmos
- Angewandte Informationstechnik Forschungsgesellschaft mbH (AIT)
- Arcanum Ltd
- Austrian Centre for Digital Humanities (ACDH)
- ArchivInForm GmbH
- Barcelona Supercomputing Center (BSC)
- Cambridge Digital Humanities
- Center for Advanced Studies, Research and Development in Sardinia (CRS4)
- Center for Art and Media Karlsruhe (ZKM)
- Cologne Center for eHumanities; Complutense University of Madrid
- Dutch knowledge centre for digital heritage and culture (DEN)
- Digitalisierung Innsbruck; DiSSCo
- Ecole des hautes études en sciences sociales (EHESS)
- Gesellschaft für Medien in der Wissenschaft (GMW)
- Institut de Recherche et d'Histoire des Textes
- Institut fur Angewandte Informatik
- Istituto Italiano di Tecnologia; Intelligent Systems Lab
- Kaunas University of Technology
- Klokan Technologies GmbH
- Knowledge Integration Ltd
- Laboratorio de Innovacion en Humanidades Digitales (LINHD)
- Laboratory on Digital Libraries and Electronic Publishing, Department of Archives
- Lexicographic Institute Miroslav Krleža
- Picturae Technische Informationsbibliothek Hannover (TIB)
- Swiss Federal Institute of Technology (ETH Zürich)
- TU Darmstadt
- TU Dortmund
- Universitat Autònoma de Barcelona
- University of Applied Sciences Western Switzerland (HES-SO)
- University of Applied Sciences in Dresden

- University of Bamberg
- University of Belgrade
- University of Helsinki
- University of Applied Sciences in Mainz (AI Mainz)
- University of Hradec Králové; University of Luxembourg
- University of South Bohemia
- Women in AI

FRAMEWORK CONDITIONS

There are some European projects on Cultural Heritage and Data Infrastructure for GLAM and the Humanities, and the Time Machine needs to be inserted into this context in order to be part of a bigger and coherent European infrastructure. Collaboration with these institutions (for example, Europeana) is already taking place.

Big developments in key areas of technology will be achieved in the next couple of years independent of the TM developments, especially in the fields of AI. The TM needs to capitalize those findings for its own goals directed at the big data of the past. This is compatible with the development of RFCs and the proposed funding mechanisms. It also allows us to leverage existing projects and funding sources in the European framework.

The TM is not only an archival infrastructure, but also a research hub and a research environment. It is important to make sure that those research results are part of a preservation and dissemination strategy based on FAIR principles. Currently there are some incipient initiatives at a national level that try to push forward these principles (Germany, Portugal, etc.) but there is a demand to make it work at a European scale. As a European LSRI, the TM can be a contributing factor to impulse proper research data management and dissemination. An important issue is licensing. Even in the Creative Commons framework the chain of attribution is hard to keep. Many sources might be proprietary. The European Commission could support TM and cultural heritage by additional legislation that e.g. links open data policies to support of cultural heritage and cultural heritage digitisation and research projects.

Every country usually has its own policy on data management. For example, the common policy in Germany is to keep all Primary data inside of the country on local servers but metadata can be shared with all partners from other countries. Data repository should be able to support selected policy and be flexible enough to switch Storage layer (Inside/Outside) or Access levels (Open/Close) if policy will change.

Since GDPR law was finally approved by the EU commission, TM repositories should be GDPR-compliant. Respective measures for anonymisation of living persons must be set into place.

RISK AND BARRIERS - MEASURES TO ADDRESS THEM

The risks and barriers for this Pillar can be divided in different categories:

1- Risks involving funding sources. The funding from some institutions might be interrupted for reasons outside of the TM's control. More generally, the funding received might not be enough to perform all the innovative and ambitious goals of the project. Thanks to the modular system of development and the Time Machine Organisation it is possible to prevent this risk up to a certain degree, as they guarantee that the project is not dependent on one particular funding source and that at least some important milestones will be met and will generate positive contributions to the field. In the worst-case scenario, the TMO can act as a hub to connect different institutions, and projects will be able to work albeit with limited funding. Yet, development and implementation of RFCs will become less likely.

2- Risks involving external technological developments. The development of some technologies in the private sector might reach breakthroughs that make the current developments in the TM slightly outdated. However, the wide network and strong links with experts and institutions throughout Europe guarantees that the TM will be up-to-date with the main technological developments at any time and will be able to adapt. The "on top funding" mechanism will enable to assimilate break-through developments to a certain degree.

3- Institutional barriers. Groups performing research in Humanities and Social Sciences have academic traditions that might not incorporate the developments and tools provided by the TM. The TM needs to get into contact with this field of research and foster innovative quantitative work on the Big Data of the Past with convincing results that make the case for the value of the TM for historical research clear. TM funding mechanisms incentivize cooperation with the TM.

Commercialisation vs. public access. Risk in used linked data services which were previously open moving to commercial models / behind paywalls. TM will link TM funding mechanisms towards public domain and open access data models.

ANNEX A: OVERVIEW OF TIME MACHINE

Rational

Over the centuries, the national, regional and local identities of Europe have evolved in relation to one another, through large swathes of transnational mobility and through dense exchanges that have shaped European languages, traditions, arts and many other aspects of human activity. These processes have largely contributed to the creation of a European culture characterised by diverse historical memories, which have laid the foundations to values and ideas harmonised by pluralistic and democratic dialogue.

To-date, however, increased globalisation, changing demographics and their threat against the idea of a shared past, as well as the resurgence of unresolved conflicts deep-seated in European memory are key drivers of a "localisation backlash" that places local and personal interests above any other. These growing trends present a clear threat to the cohesiveness of European cultural identity and sense of belonging.

Pluralistic and democratic dialogue in Europe has traditionally been facilitated by important intermediaries, such as cultural media and institutions acting as cornerstones of our shared values, principles and memories. Today, the dialogue between different actors and the historical visions they embody is complicated by the rise of private digital platforms that have created a new space of opinion-leadership, as well as new forms of political expression and participation.

Managed by proprietary algorithms, such platforms may prioritise popularity and personal agendas over historical and cultural data, opening the way to fake news. In the resulting crisis of authority that affects journalism, academia and politics, many people do not trust anymore the information received from these institutions.

These unprecedented transformations create a vital need for Europe to restore and intensify its engagement with its past as a means of facilitating an evidence-based dialogue between diverse historical memories, their values and mutual interdependencies and building a common path across generations. Time Machine responds to this need by building the required infrastructure, and an operational environment for developing the "Big Data of the Past" that will transform history and culture across Europe, opening the way for scientific and technological progress to become a powerful ally to safeguarding European identity and democratic values.

For Time Machine, digitisation is only the first step of a long series of extraction processes, including document segmentation and understanding, alignment of named entities and simulation of hypothetical spatiotemporal 4D reconstructions. The hypothesis pursed by Time Machine is that such computational models with an extended temporal horizon are key resources for developing new approaches to policy making and to offering services to European citizens and consumers.

Still, there is one more crucial reason supporting the cause of Time Machine. After the creation of the web that digitised information and knowledge and the social media that digitised people and characteristics of human behaviour, a third technology platform is being created, digitising all other aspects of our world, giving birth to a digital information "overlay" over the physical world, a "mirror-world"[96]. The mirror-world will aim to be an up-to-date model of the world

[96] The term was first coined by Yale computer scientist David Gelernter in 1991 in its book "Mirror Worlds: Or the Day Software Puts the Universe in a Shoebox...How It Will Happen and What It Will Mean" (Oxford University Press, 1991)

as it is, as it was and as it will be. All objects (including representations of landscapes) of the mirror-world will be machine-readable, and, therefore, searchable, traceable and subject to be part of simulations by powerful algorithms. In the mirror world, time will be a fourth dimension, as it will be very easy to go back to the past, at any location, reverting to a previous version kept in the log. One may also travel in the other direction, as future versions of a place can be artificially created based on all information that can be anticipated about the predictable future. Such time-trips will have an increased sense of reality, as they will be based on a full-scale representation of the present world. Time Machine is today the most advanced concrete proposal to build the first version of a European mirror-world.

Like the other two platforms, the mirror-world will disrupt most forms of human activity, as we know them today, giving birth to an unimaginable number of new ideas (and many problems) and creating new forms of prosperity from new forms of economic and social activity that will shape new behaviours and ecosystems. In this scenario that is currently unfolding, Time Machine will enable Europe to be one of the leading players, shaping the mirror-world according to its democratic values and fundamental ethics (open standards, interoperability). With Time Machine, while it will have a powerful tool to strengthen its cohesion and sense of belonging, Europe has, moreover, an opportunity to impose its own terms against the multinational technology giants that will fight for dominating this new technology platform, just as those who now govern the first two platforms have done in the past.

LSRI Structure

The Time Machine LSRI is articulated around four Pillars, each defining a specific objective:

- Pillar 1 – Science and Technology for the Big data of the Past: Addressing the scientific and technological challenges in AI, Robotics and ICT for social interaction, for developing the Big Data of the Past, while boosting these key enabling technologies in Europe.
- Pillar 2 – Time Machine Operation: Building the TM infrastructure for digitisation, processing and simulation, in order to develop a sustainable management and operational model (“TM franchise”), as well as to create the basis for and engagement with the TM communities participating in the development and use of Time Machine.
- Pillar 3 – Exploitation Avenues: Creating innovation platforms in promising application areas, by bringing together developers and users for the exploitation of scientific and technological achievements, and therefore leveraging the cultural, societal and economic impact of Time Machine.
- Pillar 4 – Outreach and innovation: Developing favourable framework conditions for the outreach to all critical target groups, and for guiding and facilitating the uptake of research results produced in the course of the LRSI.

Each Pillar comprises thematic areas, as shown in Figure A-1.

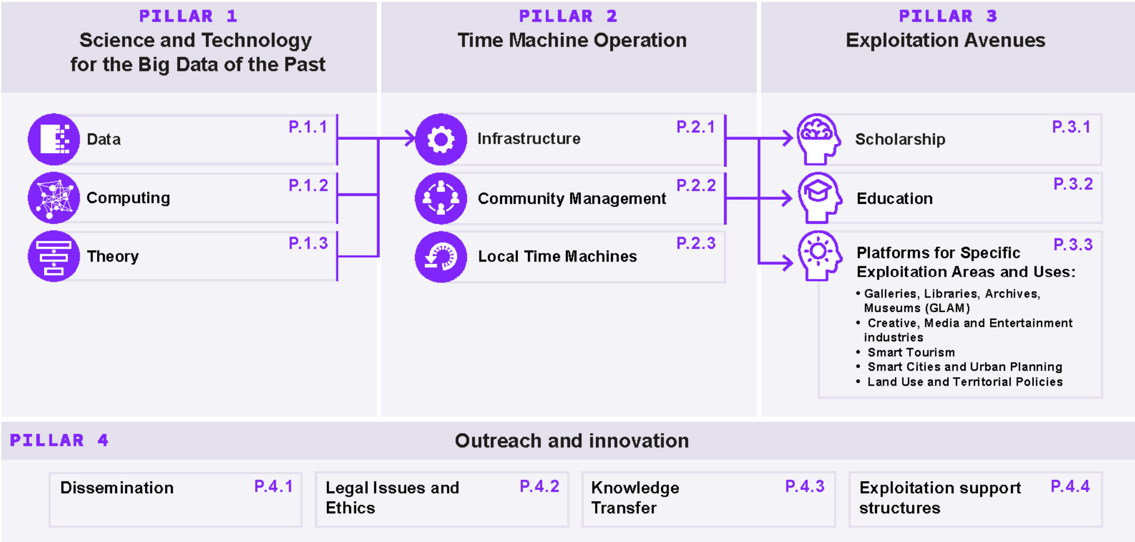

Figure A-1: Time Machine Pillars & Thematic Areas and their interrelations

Expected impact

- A strong boost in EU competitiveness in AI and ICT:
  - An AI trained on Big Data of the Past will offer a strong competitive advantage for Europe in the global AI race.
  - Disruptive technologies in machine vision, linguistic and knowledge systems, multimodal (4D) simulation, HPC and long-term data storage will strengthen the competitive position of EU industry in these fields.
- New disruptive business models in key economic sectors:
  - Cultural Heritage is a unique asset for European businesses. Time Machine will act as an economic motor for new services and products, impacting key sectors of European economy (ICT, creative industries and tourism).
  - Time Machine will develop a paradigm to follow for cities that wish to make a creative use of their historical past.
- A transformational impact on Social Sciences and Humanities (SSH):
  - With Time Machine, SSH will evolve to address bigger issues, allowing new interpretative models that can smoothly transition between the micro-analysis of single artefacts and the large-scale complex networks of European history and culture.
- Moreover, Time Machine will:
  - Be a driver of open science, as well as open (public) access to public resources.
  - Provide a constant flux of knowledge that will have a profound effect on education, encouraging reflection on long trends and sharpening critical thinking.
  - Render education for Europeans more accessible, interactive and diversified.
  - Develop new or updated legislation or guidelines in the field of AI, including ethical norms and ethical standards in areas such as access to and re-use of digital data, harmonised rules on data-sharing arrangements, especially in business-to-business

and business-to-government situations, as well as clarified concepts in data ownership.

  - Create new jobs for digital and traditional humanists and social scientists, while offering clear opportunities for talented humanities graduates with increased digital skills, by demonstrating the benefits of the new profession “Digital Humanities expert”.

- Having confirmed itself as one of the pioneers, Europe will make meaningful contributions to the foundation and use of the mirror-world, in line with its values and ethics.

ANNEX B: DEFINITIONS - ABBREVIATIONS

Definitions

4D Simulator	One of 3 TM Simulation Engines. The 4D Simulator manages a continuous spatiotemporal simulation of all possible pasts and futures that are compatible with the data. The 4D Simulator includes a multiscale hierarchical architecture for dividing space and time into discrete volumes with a unique identifier: a simulation engine for producing new datasets based on the information stored. Each possible spatiotemporal multiscale simulation corresponds to a **multidimensional representation** in the 4D computing infrastructure. When a sufficient spatiotemporal density of data is reached, it can produce a 3D representation of the place at a chosen moment in European history. In navigating the representation space, one can also navigate in alternative past and future **simulations**. **Uncertainty and incoherence are managed at each stage of the process and directly associated with the corresponding reconstructions of the past and the future.**
Big Data of the Past	A huge distributed digital information system mapping social, cultural and geographical evolution. A key objective of Time Machine is that such a system brings together dense, interoperable, standardised (linked data, preferably open) and localised (marked up with spatial-temporal information) social, cultural and geographical heritage resources.
Communities	Group of users, self-organised by territorial or transversal interests, offering various voluntary works and favours to the partners (annotation, digitisation, bibliographic research, development), according to the standards in place. These communities can elect a representative.
Digital Content Processor	Automatic process extracting information from documents (images, video, sound, etc.). Level 1 Digital Content Processor labels mentions of entities. Level 2 Digital Content Processor labels relations between entities. The Digital Content Processor of Level 3 labels rules. Each process is fully traceable and reversible. The results of the processing constitute the core dataset of the Big Data of the Past and are integrated in the TM Data Graph.
Large-Scale Inference engine	One of 3 TM Simulation Engines. The Large-Scale Inference Engine is capable of inferring the consequences of chaining any information in the database. This permits to induce new logical consequences of existing data. The Engine is used to shape and to assess the coherence of the 4D simulations based on human-understandable concepts and constraints. Its origin lies in more traditional logic-based AI technology, which has been slightly overlooked since the recent success of the deep learning architecture, but that can, nevertheless, play a key role in an initiative like TM.
Local Time Machine	Zone of higher "*rebuilding the past activities*" density. Constituted of a group of local partners and communities bound by a common territorial

	focus and a declaration of intent, which respect both graphical and values charters. Any institution that meets eligible criteria can integrate a Local Time Machine. The declaration of intent is reviewed on an annual basis (time for new partners to integrate the TM).
Project with Time Machine label (PWTML)	Project respecting the technical charter, of which tasks are documented - modelled within the Time Machine graph. All the partners of a PWTML must have signed the declaration of intent of the related Local Time Machine.
Technical Charter	The Technical Charter should contain information about infrastructure standards required within any project with the Time Machine label. The Technical Charter defines the Time Machines Rules, Recommendations, Metrics and Official software. The document is revised periodically.
Time Machine Box	Servers that allow partners to store their documents and metadata and to integrate the Time Machine Network easily and be appropriately documented in the Time Machine Graph. The Time Machine Box is part of the Time Machine Official Components.
Time Machine Data Graph	Formal representation of knowledge extracted by human or automatic process, represented with semantic web technology.
Time Machine Index	The TM index is a global system indexing different type of objects: e.g. documents; iconography; 3D geometries. It gathers all information regarding documents and their contents and could be used as a basis for other search engine infrastructures (it allows backups).
Time Machine Infrastructure Alliance	Coalition of TM's partners regrouping in-kind donators for infrastructure components (server's space and computing power).
Time Machine Mirror World	One of the API of the Time Machine using the processing of the 3 TM Simulation Engines to produce a continuous representation model that can be accessed as an information stratum overlaying the real world.
Time Machine Network	Set of all the partners *actually* interacting in the Time Machine. Each member of the Time Machine Network must have signed the Value and Technical Charter.
Time Machine Official Components	Pieces of software (e.g. Time Machine Box) that help partners conforming to the Time Machine rules seeing as they are directly embedded in the software.
Time Machine Operation Graph	Formal representation of the past, on-going and future operations of the partners in the Time Machine Network and the data pipelines.
Time Machine Organisation	Association regrouping the Time Machine Partners, active or not. Not all may have signed the Values and Technical Charters.
Time Machine Recommendations	Recommendations on technology which are not obligatory at this stage for the development of the Time Machine (e.g. choice of a particular IIIF image server).
Time Machine Request for Comments	Main document for the progressive design of the Time Machine infrastructures, standards, recommendations and rules, inspired by the process used for 50 years for the development of Internet Technology, today administrated by the Internet Engineering Task Force (IETF) as part of Internet Society (ISOC).
Time Machine	Standards and rules that need to be followed to be acceptable in the

Rules	Time Machine Network and become a Time Machine operator. Any entities not following these rules are out.
Time Machine Standard Contracts	Set of standard contracts to facilitate the interaction between Time Machine partners.
Time Machine Standard Metrics	Measures helping partners of the Time Machine Network coordinate with one another to compare performance (not only for quotes of services, but also for research performances, etc.).
Time Machine Super Computing Architecture and Simulation Engines	TM Super Computing Architecture composed of distributed computing resources from the TM Network provided by the TM Infrastructure Alliance. On this distributed architecture, different typologies of computing processes can run. For instance, Digital Content Processors are intrinsically easier to run in parallel, whereas Simulation engines, which allow users to generate possible pasts and futures from the TM Data Graph need for more specific computing architecture.
Universal Representation Engine	One of 3 TM Simulation Engines. The Universal Representation Engine manages a multidimensional representation space resulting from the integration of the pattern of extremely diverse types of digital cultural artefacts (text, images, videos, 3D), and permitting new types of data generation based on transmodal pattern understanding. In such a space, the surface structure of any complex cultural artefact, landscape or situation is seen as a point in a multidimensional vector space. On this basis, it could generate a statue or a building, produce a piece of music or a painting, based only on its description, geographical origins and age.
Values Charter	Conform to the principle of openness in EU law.

List of abbreviations

AI	Artificial Intelligence
CH	Cultural Heritage
GLAM	Galleries, Libraries, Archives, Museums
LTM	Local Time Machine
PWTML	Project with Time Machine Label
RFC	Request for Comments
SSH	Social Sciences and Humanities
TM	Time Machine
TMO	Time Machine Organisation

ANNEX C: RESPONSES TO PROJECT REVIEW REPORT

Comment	Response
Recommendations to bring the final version up to the expected quality level	
a. Focus on clarity, conciseness and quality of the text, rather than quantity and length. The text is often verbose and some passages are unclear and the rigorously applied structure impedes conciseness and readability.	The project would be willing to address specific comments rather than general impressions that might relate to non-objective qualifications of writing style.
b. Give more emphasis to the concept of multiple 'Mirror Worlds'. If credible and veracious reconstructions of the past and plausible predictions and models of the future can be made with the help of the available data, that would already be a ground-breaking result.	We now clarify that Mirror Worlds is a likely technology to be established within the context of Time Machine. Furthermore, we now emphasize the importance of causal models that are key towards the estimation of unseen data. As such, we follow the approach presented by Pearl et al. as now detailed in the roadmap description.
c. Given that this is a road map, more concrete actions should be described. There is very little indication of what will actually be done, who will do it, and who will pay. The concrete intentions in the original proposal ("We will develop solutions for ...") have become diluted into more repetitions of the vision. So it is essential to provide more concrete statements of the specific steps and processes envisaged to proceed from the state of the art to the implementation of the vision.	Concrete actions are the calls for projects that are continuously synchronised using the TM RFC system. As such, we are inherently able to adopt to new and unforeseen technological developments and breakthroughs. We regard a fixed waterfall model of development steps as too risky for such a large-scale project and thus decided for a more agile development path.
d. Many of the most exciting possibilities suggested in the vision (e.g. the concept of multiple 'Mirror Worlds'; scanning boxes of paper without opening them; or the 3D recreation of a site that reconfigures as a slider moves down the centuries) should receive further focus as they are only scarcely mentioned.	There is a multitude of prior work towards any of the mentioned examples in the consortium. We deem all of the technologies within the roadmap as important. Highlighting only few examples might appear to introduce a bias towards those and will misrepresent the broad coverage of technologies within the consortium.
e. Some early building blocks ought to be ready in this preparatory phase, e.g. the draft charters (graphical and values and technical); the prioritisation strategy.	The project considers that what is termed as "building blocks" should be treated by RFCs to be developed as part of the roadmap implementation.
Additional detailed comments	
a. The deliverable identifies methodologies and tools, e.g. metadata formats, reference models, and vocabularies, that are either cumbersome in use because the implementation is very labour intensive (e.g. CIDOC-CRM, Dewey Decimal Classification), or limited in practice by IPR issues (e.g. Iconclass, Getty Vocabularies), without clearly describing how innovative technologies or legal models may overcome the limitations they entail.	We now address this issue in more detail. Please note that Pillar 1 has a focus on science and technology. Legal topics are dealt with in Pillar 4.

Comment	Response
b. The discussion often glosses over challenges at the granular level (e.g. rights issues in film; cross domain indexing). Issues around multilinguality are poorly addressed. There are few examples of current leading edge initiatives and no estimates of volumes at stake (e.g. institutions, items to be digitised, sites to be surveyed). It is very hard to follow the thread from the overview to the milestones.	The issue of multilingualism is now addressed more extensively.
c. There is a suggestion (Milestone 1.2.2, page 29 on screen pagination) that data already in distributed storage systems will be copied. Why? Duplicate datasets get out of step very quickly. And, surely, a single data warehouse is not envisaged; the scale of that is clearly prohibitive, and suggests that the architecture for aggregating and searching distributed datasets is under-developed.	Copied data will be versioned in order to keep track of possible changes. Distributed storage systems have the disadvantage that they may suffer partial or complete damages as such cultural heritage of high priority needs to be backed up with several decentralized copies in order to prevent loss of data. Data back-up and long-term storage is not to be confused with distributed aggregation and searching structures. A distinction is made between backup (which involves duplications) not visible to the end-user, and what is visible to the end-user for which there is only one source at a given time (and versions of that source during time).
d. The discussion in Section 3, Social Sciences and Humanities describes how users use the assets, but adds little value. It is a slightly more detailed version of what was in the DoA, and is more of a wish list of potential impacts.	We strongly believe that we have to implement the use of Time Machine in applications to show immediate benefits. Within the short time frame of the CSA phase this is impossible to achieve for the full project. Nonetheless various Time Machine members already drive such efforts ahead within the scope of Local Time Machines as detailed in Pillar 2. Yet, we are only at the beginning of a major large-scale effort and only Time Machine will enable appropriate synchronisation of the various regional efforts at present