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

Deliverable D2.1
Science and Technology (Pillar 1) Roadmap-draft

Abstract

A draft 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	D3.1
Deliverable Title	Science and Technology (Pillar 1) Roadmap - draft
Work Package	WP2
Delivery Due Date	30 June 2019 (Month 4)
Actual Submission Date	1 July 2019
Leading Partner	FAU, UA
Main Author	Andreas Maier, Gustavo Fernández Riva, Mike Kestemont
Contributions	Alicia Fornes, Björn Eskofier, Bruno Vallet, Ellen van Noort, Fabio Vitali, Fauzia Albertin, Florian Niebling, Francisco Casacubierta, Gernot Fink, Guillaume Touya, Hanan Salam, Jacqueline Klusik-Eckert, Jeroen Deploige, Joan Andreu Sánchez, Lars Wieneke, Marieke van Erp, Michelle Lindlar, Peter Bell, Robert Sablating, Sander Münster, Nanne van Noord, Thierry Poibeau, Thomas Bähr, Tomasz Parkola
Reviewer(s)	Frederic Kaplan

History of Changes

Date	Version	Author	Comments
28/06/19	1.0	Andreas Maier, Mike Kestemont, Gustavo Fernández Riva.	Document ready for submission, after integration of comments by internal reviewers.

Disclaimer

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 the social, cultural and geographical evolution. A key objective of Time Machine is that such 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 (annotation, digitisation, bibliographic research, development), according to the standards in place, to the partners. Those communities can elect a representative.
Digital Content Processor	Automatic process extracting information from documents (images, video, sound, etc.). Digital Content Processor of Level 1 just label mentions of entities. Digital Content Processor of Level 2 label relations between entities. Digital Content Processor of Level 3 label Rules. Each processing 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 Large-Scale Inference Engine is used to shape and to assess the coherence of the 4D simulations based on human-understandable concepts and constraints. Its origin comes from more traditional logic-based AI technology, slightly overlooked since the recent success of the deep learning architecture, 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 who meets eligible criteria can integrate a Local Time Machine. The declaration of intent is reviewed on an annually basis (time for new partners to integrate the TM)
Project with Time	Project respecting the technical charter, whose tasks are documented -

Machine label (PWTML)	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	Should contain information about infrastructure standards required within any project with 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 integrate easily the Time Machine Network 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. Could be used as a basis for other search engine infrastructures (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 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 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. Some maybe active and other not. Not all may have signed the Values and Technical Charters.
Time Machine Recommendations	Recommendation 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 Rules	Standard and rules that need to be followed to be acceptable in the Time Machine Network and become a Time Machine operators. Any entity 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 (for quotes of services, but

	not only, there are also use 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 process 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

Table of Contents

INTRODUCTION

Time Machine (TM) is a Large-Scale Research Initiative (LRSI), pushing the frontiers of scientific research in Information and Communication Technologies (ICT), Artificial Intelligence (AI) and the Social Sciences and Humanities (SSH).

TM is built around the vision to develop the Big Data of the Past, a huge distributed digital information system mapping the European social, cultural and geographical evolution. This large-scale digitisation and computing infrastructure will enable Europe to turn its long history, as well as its multilingualism and multiculturalism, into a living social and economic resource for co-creating a common future. The proposed LRSI will use space and time as shared references across domains, disciplines and cultures, to understand and give value to constructions, artefacts, observations and data produced over centuries, enabling Europeans to better appropriate their heritage and strengthen the feeling of European belonging.

The key objective of the TM CSA project is to develop a full LSRI proposal around this TM vision. Detailed roadmaps will be prepared, organised around four pillars, namely science and technology, TM operation, exploitation avenues and framework conditions. The roadmap development methodology foresees the elaboration of draft roadmaps for each pillar by working groups composed of Consortium experts, followed by a round of consultations with relevant external stakeholders. These consultations will enable the Consortium to finalise the pillar roadmaps in a way that reflects the needs and expectations of a pan-European ecosystem that has been built around Time Machine and is currently expanding at fast rate.

The roadmap for the science and technology pillar is developed in WP2. This document is the formal deliverable D2.1 presenting the draft roadmap for this Pillar 1 and details the broader vision underlying it. The emphasis is on describing the qualitative aspects of the proposed research and innovation actions in a sufficient level of detail, enabling informed feedback to be received during the consultations that will follow. The final roadmap is planned for Month 8 (October 2019).

Following this short introduction, the deliverable is organised in sections presenting:

- An overview of the TM LSRI and the driving lines for the design of the science and technology pillar.
- The research and innovation plans for pillar 1, detailing the state of the art, the targeted achievements and the methodologies to obtain them.
- the funding resources that can support the pillar 1 actions.
- the stakeholders to be involved in and/or that are directly concerned by these actions.
- the framework conditions that relate to the implementation of pillar 1.
- the risks and barriers related to Pillar 1 and the strategies foreseen to address them.

DESIGN OF PILLAR 1 – SCIENCE & TECHNOLOGY

Overview of the Time Machine LSRI

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"[1]. The mirror-world will aim to be an up-to-date model of the world 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.

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.

[1] 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)

- 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.

LSRI Structure

The Time Machine LSRI is articulated around four pillars, each defining a specific objective of the initiative:

- 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 1.

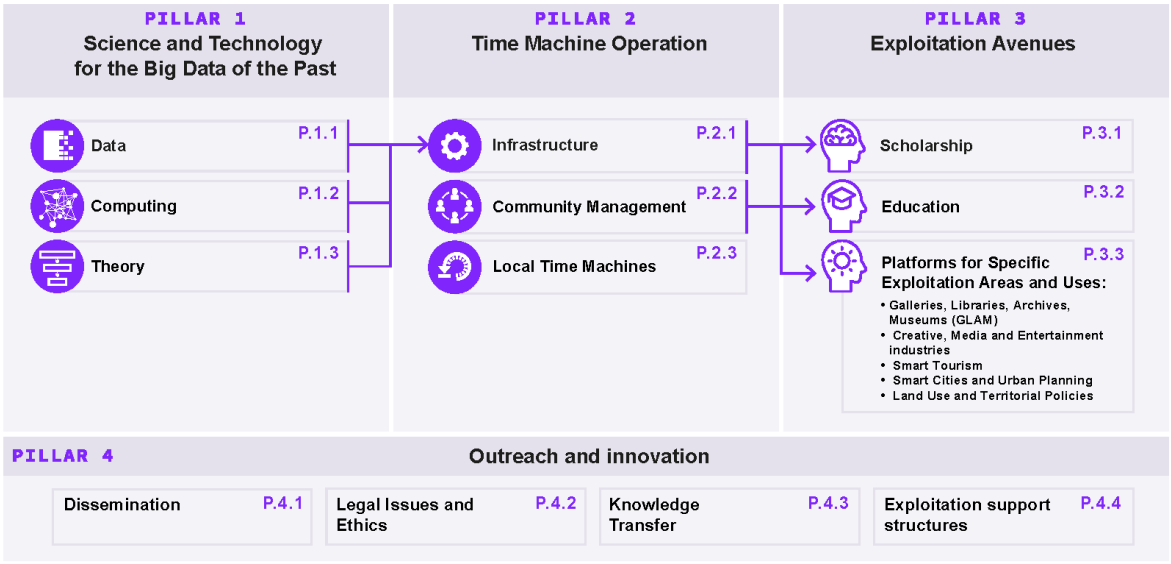

Figure 1: Time Machine Pillars & Thematic Areas and their interrelations

Time Machine will be designed in four phases:

- Bootstrapping – 2021-2023
- Scaling – 2023-2025
- Sustaining – 2025-2027
- Globalising – 2027-2030

Pillar 1 approach

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 on its 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 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 its own disciplinary traditions, methodologies and long-and mid-term goals. In order to implement and shape the developments outlined in this document, 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. The discussions and results of work carried out in Pillar 1 will in most cases 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 traveling by extracting knowledge and establishing links over space and time. We aim to put together multidisciplinary work groups in Europe 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 development of the roadmap 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 detailed in the next section on the actual roadmap, where a series of realistic milestones are proposed to be solved via 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 for Pillar 2, because the targeted achievements in Pillar 1 directly feed the exploitation plans detailed there. 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 realise hotspots of local digitisation initiates, 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 realise a distributed super-computing and storage system, which can be efficiently indexed, searched, updated and exploited?*

The implementation of the Pillar 1 roadmap will require the involvement and coordination of many actors across the EU, which defines further requirements for the governance scheme of the initiative, to be designed in WP6.

The whole governance of Time Machine is based on a **Time Machine Organisation (TMO)** that sets the global rules for all actions and operations related to Time Machine, including definition of processes, labelling system and recognised infrastructure. In this context, coordination has two aspects:

- The need for a clear strategy to organise the research link of different initiatives at European/national/regional scale with Time Machine research objectives. This means that boundaries may have to be drawn between official Time Machine research and research

that contributes to the Time Machine goals but in a more indirect manner. There are different possibilities including organising “Time Machine Calls” and creating a “labelling system” for ongoing projects (financed by other national, philanthropic or European funds).

- The need for a monitoring strategy to track progress inside the TM official research projects but also the global progress of state of the art matching the Time Machine Research Objectives that gives a feedback in the broader TM strategy.

The development of the governance scheme should take into account these requirements, by designing appropriate organisational structures and processes for the formulation and implementation of the TM strategy to be followed in the different phases, as well as an efficient monitoring system that ensures that overall progress in on-track and that lessons learned over an implementation phase are used in programming subsequent phases.

RESEARCH AND INNOVATION PLANS

State of the Art

To realise 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 roadmaps, 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 emphasises 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 digitise and model CH Objects which in Time Machine is extended to cities and territories. 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.

In **(1.1.1) 2D Digitisation** several programs exist, produced at different times and places, with a heterogeneous coverage of objects. Digitisation benches equipped with digital camera backs are the main technique for the task at large scale and, although they vary greatly, most tend to be manually operated or at least manually assisted. Traditional flat-bed scanning is used at smaller scale. The technology for high quality digitisation got 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 Metamorphose guidelines, IMPACT Centre of Competence recommendations for digitisation projects and the FADGI of Library of congress (USA).

(1.1.2) 3D Digitisation 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. Moreover, the typology of 3D digitisable objects are really varied in terms of scaling: landscapes, archaeological sites, architectures, sculpture, paintings, books, etc. Multiple methods have been developed, although none of them has succeeded yet to become a routine procedure, especially for historically important objects. A list of the currently available 3D acquisition techniques includes structured light, stereo vision, digital holography, photogrammetry, many different kinds of spectral imaging, as well as optical coherence tomography, X-ray tomography, ultrasound, Lidar, etc.

2D and 3D digitisation are essential for the TM, as most of the relevant objects in the Big Data of the Past can be included in these categories. In particular, fast and large-scale acquisition is an important goal in TM. These 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 are of high importance in the coming years. However, when considering the more recent history **(1.1.3.) Audio, (1.1.4) Film, and Video** digitisation are also relevant: they have each a set of very specific problems and techniques. Video and audio carriers are deteriorating rapidly and the original playback equipment needed to do the digitisation is getting obsolete. Film digitisation is very labour intensive and its very likely that in the upcoming years new technologies will be developed. Automated metadata acquisition, like speech to text, is therefore also expected to be improved in the near future.

Apart from traditional CH, like documents, paintings, sculptures, and buildings, the TM will also digitise bigger geographical features like cities, landscapes, and territories, as well as other environmental information like 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, statistical surveys into different products like digital elevation models, topographic databases, land use products, and virtual reality models. These kinds of objects pose a particular challenge for digitisation, which current technology is only inadequately addressing.

Finally, data acquisition is not limited to creating digital images, but also extracting some of their invisible physical and chemical features, what is known as **(1.1.5) Scientific Analysis**,

which is extremely important to ensure correct preservation and restoration of objects as well as the sustainable management of our environment and places to live.

1.2. DATA MODELING describes the conceptual framework, the technological approaches and the knowledge representation tools necessary to support and harmonise the digital forms of cultural artefacts in order to allow their fruitful interoperation with each other. This means handling primary, secondary, and tertiary data as well as disputed 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). In addition to that, research infrastructure should make it possible to interconnect these data with the use of Linked Open Data technologies and appropriate semantic enrichment. All data should be stored in these FAIR repositories with possibility to get data in and out in the standardised way. It will allow to make data processing and transmission pipelines transparent and migrate data from one repository to another without any difficulties and following specific policy.

The field of **(1.2.1) Knowledge Modelling** deals with uncertainty and trust, provenance, layered annotations, graph models, and multilingual data. These are complex tasks, as secondary and tertiary data over the same artefact are intrinsically multiple in number and often inconsistent or contrasting in content. This is, however, not meant as a limitation but is intrinsically part of the right way to handle and debate cultural assets. CIDOC-CRM and FRBRoo are the current state of the art to deal with this issue. Uncertainty needs to be assumed for every piece of information, and trust must be asserted based on agreement, reputation of the source and relevance of the assertion. Geographical data is no exception to these problems, and is characterised by heterogeneity, differences of conceptual models used by humans to model the real world into objects or into fields, differences of implementation of the spatial aspect of the model into vector or raster data, differences in resolution and in accuracy related also to the technologies available. They are also characterised by implicit information carried by geometries; not all spatial relationships are explicitly represented, rather they can be inferred from algorithms applied to geometries. In the past twenty years, much effort has been devoted to geodata interoperability by designing metadata standards, with a specific attention to ontologies that are required to align different conceptual models. This area also will also consider workflows and models that respect the correct handling of rights, although that issue will be addressed in more detail by Pillar 2.

In the TM model, every piece of information will have to be associated to temporal and geographical characterisation. The existing OWL-Time ontology as well as the Time-Indexed Value in Context design patterns are a good start, but more flexible Data Modelling also includes work on different **(1.2.2) Data Formats**, which are strictly media-dependent and should follow the most recent best practices of the data type they represent, be non-proprietary, feature rich and widely used. XML for texts, TIFF and JP2000 for pixel or voxel-based images, MXF D10-50, JP2000, AVC-intra100, Quicktime prores, ffv1, DPX, XDCam HD, H264 for video and film, Wave, MP3, flac for audio, STL and COLLADA for 3D scans are appropriate for primary data, while RDF is appropriate for secondary and tertiary structured data, as well as XML for secondary textual data.

The concrete data model and ontologies for primary, secondary, and tertiary data will be one of the most relevant and crucial aspects of the initial phases of the Time Machine project, and widespread and correct adherence to such model will be instrumental to the success of the overall LSRI. Due to the very different kind of CH objects and data formats, it is important to develop appropriate **(1.2.3) Metadata Formats and Mapping between Standards**. The TM should be able to incorporate and link to already existing metadata element sets (not only DC,

SKOS, and the CIDOC CRM, but also e.g., EAD for archives, METS for digital libraries, LIDO for museums, TEI for literary texts) and from the vocabularies of values (authority lists, thesauri or controlled vocabularies) used by data providers (such as Geonames, Getty Vocabularies, Iconclass, Dewey Decimal Classification, DBPedia.).

The Time Machine should offer a user interface for **(1.2.4) annotation** to offer to the LTMs. This way, the standardisation is easier to guarantee, as the rules will be embedded in the platform. This interface will be created following current knowledge of best practices in human-computer interaction and use AI to assist in the annotation (cf. Task 2.4 in the taxonomy).

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

1.3. LONG-TERM PRESERVATION describes the sum of processes undertaken to maintain a digital object's availability and interpretability – from a technological as well as a semantic perspective: the digital preservation. Risks can be classified as falling into different classes.

The **(1.3.1) Bitstream Layer** (Storage Layer) shall mitigate technical (e.g., bitflip, media failure), human (e.g., accidental deletion/manipulation) and environmental (e.g., fire) risks. Best practice is 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 utilisation of cloud storage providers grows and data integrity is verified via checksum algorithms which are typically stored in local, centralised databases. There has 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 **(1.3.2) Functional Layer** 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 not fulfil user requirements anymore.

The **(1.3.3) Semantic Layer** deals with risks associated with the context or knowledge loss such as in form of missing contextual information to interpret information within an object in the first place or semantic change over the course of time. The TM should adhere to the requisites of **(1.3.4) Trustworthy Archives** proven through certification measures such as CoreTrustSeal, DIN 31644 nestor Seal or ISO 16363 certification processes, ISO 14721:2012 Open Archival Information System Standard and be vigilant to the developments, like the forthcoming new ISO 14271:2019 version, now also including an outer-OAIS/inner-OAIS model, where functional entities are spread across different organisations. The goal of the TM in this area is to create guidelines, best practices and recommendation, as well as a framework and 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 public 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 already mean a

significant contribution to Europe's engagement with its CH. Nevertheless, it only concerns one part of what 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 initiative to improve the exploitation and understanding of CH objects. Especially in the field of Machine Learning, these will dramatically improve Europe's visibility and contribution to global research challenges. The developments in these fields are essential for the creation of the TM Official Components. The work on Computing and AI is divided in the following areas:

2.1. COMPUTER VISION AND PATTERN RECOGNITION deals with ways for computers to transform data into valuable information that can be processed and analysed.

One of the most important subfields is **(2.1.1) Text Recognition**, as a big part of the sources of the Big Data of the Past 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, and neural network models have shown success for handwritten texts in favourable situations. For historical texts this is often not the case. 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 chance to train fully scalable, robust models for any kind of collection as well as high accuracy on heterogeneous documents. TM will generate open access tools able to perform these tasks.

However, not all documents use text, some of them like maps, music scores and technical drawings communicate information in their own language and alphabets. The field of **(2.1.2) Graphic Document Processing** will develop methods to process this kind of documents. Printed mathematical and musical notation already show good results, but handwritten ones are not satisfactory and much research in needed. On the other hand, maps and architectural drawings stand out as a difficult but important problem that could be solved within the framework of the TM. Currently, there are only isolated, map-type oriented image processing algorithms, 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.

(2.1.3) Image Processing and Analysis is closely related to the two previous topics and researches general computer processing of old documents needed for many other tasks, like 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 are Generative Adversarial Networks (GANs), still mainly applied to contemporary content. 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.). 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 and for many tasks, while incorporating side-information from historical sources. For example, given information on the function of a building and images from the neighbourhood, a better reconstruction of the façade of a building is likely.

(2.1.4) Indexing and Retrieval deals with methods that make searching the content efficient and successful. With the amount of data TM will generate, this is a crucial area. Currently, the cutting-edge solution is Probabilistic Indexing, 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 at super large scale as well as searching engines completely operative based on probabilistic indexing and capable of operating for any kind of cross domain content.

The field of **(2.1.5) Understanding and Interpretation** enables reasoning over visual content, creating tools to automatically gain information about semantics, style, dating, and location of a document. It also includes video and image captioning. This is a domain where major breakthroughs should be expected in the upcoming years, but not necessarily focused on cultural heritage. The TM should, on the one side, redirect those new technologies to this field and, on the other, actively contribute to the area taking advantage of the materials collected and the expertise of its members. Particularly, being able to infer geographical information will be a strong component in the TM to improve the simulation engines.

(2.1.6) Recognition and Detection focuses on the problem of identifying entities (persons, animals or objects) in different kinds of images (photographs, paintings, films, video, 3D). Currently, there is good performance for data-rich, single-modality recognition and detection of prominent classes for contemporary content, with deep learning-based techniques. 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 relevant problem of recognition and detection is **(2.1.7) Person and Face Identification**, where the gender, race and age bias in the training materials of current models is an important issue to address, as well as the need to develop models that can work with limited training data from historical sources, which are derived from heterogeneous objects including statues, paintings or coins.

(2.1.8) Modelling, Registration, and Reconstruction is about 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 available. Automated tools for national mapping agencies exist for the generalisation of topographic data and maps, but these often do not allow real-time operation. Automatic georeferencing and spatial conflation currently 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 and remote sensing imagery and the derived reconstruction results (including a versioning system).

Finally, the TM will also work with **(2.1.9) Audio Recognition and Transcription**. 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.

2.2. NATURAL LANGUAGE PROCESSING (NLP) is of the utmost importance for the TM as many of the documents to be digitised include text in a natural language. This field includes some of the most important and revolutionary innovations in the past and upcoming years as machines get exponentially better at understanding human language. However, most of those innovations

are aimed primarily at the leading modern languages and in particular English. The TM will be in charge of using these revolutionary technologies for older language variants for as many European languages as possible. 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 for 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 ones 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 task in this domain is **(2.2.1) Methods for Resource-Scarce Languages**, which aims 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 **(2.2.2) Orthographic Normalisation and Variation Handling**, as older documents tend to be written with different orthographic conventions as modern ones and many NLP tools are optimised to contemporary spelling and text conventions.

The third core topic in this domain is **(2.2.3) Machine Reading / Document Understanding / Question Answering** 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 its meaning. The current state-of-the-art is not robust against text and genre variations. Two further 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 **(2.2.4) Metadata Extraction, Manipulation, and Translation / Mapping** is devised. The research in 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 will work closely with the team in charge of task 1.2 "Data Modelling". **(2.2.5) Discourse analysis** 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 time of composition is the main issue to tackle, as most available tools were designed for modern language variants and the TM will expand the current scope to the Big Data of the Past. Machine translation can be a very important aspect of the TM, as it deals with many language variants, as also as 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 be able not only to offer reliable translations of older languages variants into the corresponding modern language, but also to other European languages.

2.3. MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE is an autonomous field of development for the project and also a transversal discipline that will influence many other areas.

(2.3.1) General Artificial Intelligence aims at solving tasks that are usually considered

to require human intelligence. In particular, strong AI that is able to perform human-level reasoning is unknown. TM might bring significant advances in this field due to the large amount and variety of annotated data that will be available in the project. One promising first step in this direction is meta learning that aims at learning of model discovery and re-use.

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 **(2.3.2) Supervised Learning** 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 **(2.3.3) Unsupervised Learning** only the data set but no label information is available. Consequently, unsupervised learning can identify hidden structure in data. In **(2.3.4) Weakly Supervised Learning** situations are considered where annotated data is scarce or not available for the target domain at all, like where 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 **(2.3.5) Transfer Learning** 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 heterogeneous and the quality of the existent annotation may vary widely.

Many of the recent breakthroughs in Machine Learning have been made using artificial neural networks, which has led to the emergence of a new area called **(2.3.6) Deep Learning**. Models in this area set themselves apart from classical models by ramping up model complexity considerably. Incorporating and contributing to 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 **(2.3.7) Universal Representation Space** that is able to describe meaning and semantics of objects, text, images, and other information sources and is able to transfer all of these representations into each other extending on recent progress in machine translation is a major aim in TM.

For any machine learning framework or method one can ask meta-questions. In this regard, **(2.3.8) Explainability** of models is a desideratum, if predictions are to be accepted by domain experts. It is also needed to understand failure of the model and identify how to address these cases. In particular fusion with prior knowledge is a promising technique towards this goal.

While all data-driven approaches rely on training data, it is still an open question how to minimise unwanted implications of the composition, selection and availability of training material in order to allow algorithms and applications to be fair and unbiased, which should be addressed as part of the topic **(2.3.9) Bias and Fairness**, which is very 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 is **(2.4.1) User-cantered interfaces**, which deals with the creation models and interfaces for end-users. This 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 mostly to the end-user interface, which enables the interaction with the Big Data of the Past. **(2.4.2) Access to large-scale information retrieval and recommender systems** deals with the ways in which user queries will be processed for such a huge database and how the information of searches will be harvest 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 are poorly used jointly. Deep learning techniques just begin to manage them jointly with a mitigated efficiency, for example by generating missing modalities in the input to query a multimodal dataset.

In the field of **(2.4.3) virtual, augmented, and mixed reality**, there are currently multiple technologies (HMDs, CAVEs, Powerwalls, L-shapes, etc.) 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 focusing on how 3D and 4D models of cultural heritage objects, buildings and sites can be incorporated in a virtual, augmented, or mixed reality environments for GLAM, education, and games.

(2.4.4) Accessibility and learning, adaptive and cognitive interfaces deal with user interfaces that change to adapt as best as possible to the needs of users and are optimised for a learning experience. 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.

(2.4.5) Motivational design (gamification, storytelling) is a design strategy to increase motivation and participation, in other words, motivate visitors to interact with places of interest and support a deeper immersion. Elements such as badges or rankings can evoke strong positive as well as negative emotions in the target group, competition-promoting approaches can be experienced as motivating and challenging and the attitudes and expectations of the target group should always be inquired about in advance when selecting gamification elements.

(2.4.6) Big data visualisation deals with specific problems of transforming big amounts of data into graphs, plots, maps and other kind of visualisations that enable the human observer to understand patterns and draw reasonable conclusions. It includes also real-time in-situ visualisation, which demands the development of new methods to be efficient. At the same time, 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 and also ways of representing metadata in an efficient way.

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

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 depending on 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 as motion sickness or plausibility.

(2.4.8) Virtual Research Environment is based on digital twins, a digital representation that mirrors a real-life object, process or system. They are robust models, linked to the real world and driven by AI, which enable interaction with them 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 creates the necessary technology to display faithful and high quality images. These are very important for the TM as 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 network of experts in computing.

(2.5.1) Rendering is the main issue, the automatic process of generating an image from a 2D or 3D model. Currently, there are high quality offline rendering to reconstruct the impression of artefacts in their original environment; real time rendering of reconstructed sites even 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.

(2.5.2) Animation is about the ability to create virtual images in movement. Currently, motion capturing using specialised suits and fine facial capture using a large set of facial markers are ideal to create models for individuals. At the same time, AI generates convincible character animation and crowds.

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

(2.5.4) Interactive Computer Graphics and Computer Games studies ways to engage users with the computer-generated images by giving them agency, the technological counterpart of topic 2.4.5. 'Motivational Design'**.**

Another related problem is **(2.5.6) Procedural Content Generation**, also known as open world generation, which is especially relevant for big simulations of historical cities. Currently, shape Grammars/Procedural Descriptions by example 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 needs to be aware of the existing technologies and infrastructures and develop methodologies if necessary.

(2.6.1) Scaling and distribution will keep track on important developments on large-scale and distributed computers. Currently 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. This requires a software redesign in order to adapt to the new hardware. First Exascale computing probably will not be a general-purpose machine and will not be available for the TM project, but the architecture of the Exascale machines can be a good example of technologies offered by future computing centres.

(2.6.2) Dynamic provision of computing platforms deals with the more concrete problem of creating computing platforms for the TM. Computing centres should set-up computing platform 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 are the next two fields. **(2.6.3) Cloud computing** investigates if this currently widely used technology could be applied for the TM, considering its advantages and potential problems. **(2.6.4) Secure distributed computing** 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 these 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 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 **(3.1.1) Qualitative and Quantitative SSH** interrogates the apparent clash of qualitative and quantitative research methods in the humanities, revised particularly with regard to an increasing plurality of approaches. The quantitative methods still encounter resistance in the midst of the traditional academic community in SSH for many reasons: the failure of previous approaches in a time without enough reliable data, resilient scholarly practices and lack of clear perspectives of the advantages of qualitative methods. The TM will try to change this state of affairs, by showing how the Big Data of the Past can have a positive influence on history related research, creating acceptance for different scholarly cultures and publications types, establish best practice guidelines for traditional scholars about born digital sources, professionalization of Digital Humanities regarding a more critical analysis, emphasis on Humanities Performance in interdisciplinary projects. The aim would be to theoretically and practically overcome the divide between quantitative and qualitative approaches, in order to provide the conceptual and

methodological framework for SSH scholarship which can combine the strengths of the tradition 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 advantages of quantitative methods (seeing patterns in large datasets and analysing those with statistical methods).

(3.1.2) Increased research scope in SSH focuses specifically in the ways in which the access to the Big Data of the Past can improve research in SSH. Traditional disciplines like history and philology have worked primarily with a methodology that induces general principles based on a limited number of sources and observations. The TM offers the possibility to expand exponentially the sources available. However, those disciplines still need to develop strategies to handle this new availability of digitised materials. The TM needs to actively contribute to create this paradigm shift and not only provide 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. 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 methodologically, 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 a situation where scholars can research on their traditional research topics more easily than ever, choose their scale of reading, play with multiscale readings, cross different kinds of sources, choose the timespan of their research, be able to extrapolate some information if needed and explore previously unknown sources.

One of the areas where the TM will try to cause a deep impact with innovative research methods is **(3.1.3) Simulation Studies**. 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. Simulation enables testing theoretical models (hypotheses) in cases where direct experiments on the real system are not possible due to practical constrains 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. This is an incipient field, with very promising possibilities, where 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, a fully integrated platform for the interface between simulation and data including model development, calibration, validation, documentation, and dissemination through a standardised protocol and extensive documentation and training material to encourage reuse and replication.

(3.1.4) Digital Methods 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, reflecting the whole life-cycle of a humanities research process: from the development of a research question, to information retrieval, source criticism, analysis and interpretation, to developing an argument and producing a narration. It thus comprises a set of skills: infrastructure criticism, algorithmic criticism, data criticism, tool criticism, interface

criticism. It also interrogates the innovations allowed by 'distant reading': 1- they can be used heuristically, whereby the patterns observed lead to new hypotheses on the phenomenon under investigation, that then subsequently are 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. **(3.2.1) History** will be at most influenced by the TM, as the data uncovered will enable the application of research methods previously only partly possible, which focus on large amount of quantitative information. Although computational methods in historical research go back to the '70s and 80's, they are still underdeveloped and mostly marginal. 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 the mere fact of 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 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 currently hard to analyse by hand (such as statistical relations between socio-demographic factors and other data on e.g. migration, language use, cultural consumption, or voting behaviour, etc.).

In **(3.2.2) Language and Literature** 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. The distant reading in literature, that usually is only available for the 19th century, 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 at new holistic perspectives on literary texts, which are contextualised 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, Establishment of novel paradigm for corpus-based research, which enables systematic quantitative-qualitative analysis especially of specialised and historical corpora, achieving statistical significance with far smaller data sets than purely quantitative techniques.

In **(3.2.3) Archaeology,** the TM might also be a great contribution. In the wake of the 'Malta' European legislation (protection archaeological heritage), enormous expansion of digital archaeological data in the past decade, mostly in the framework of 'commercial archaeology', producing an enormous mass of so-called grey literature. National data-infrastructures are prevalent, with different terminologies, data standards, and access rules. 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 a big amount of data from at least the urban areas of Europe as well as innovative 3D models and visualisations.

In the fields of **(3.2.4) Art History and Media Studies,** digitisation is less advanced than

in archaeology. On the one hand, many databases have been developed in the past, but with very heterogeneous models and mostly proprietary software. On the other hand, many institutions have digitised their holdings, although focusing on the most famous objects. Some projects, like Europeana have improved the situation by providing a framework for European art digitisation projects. 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 many ways. First, for GLAM institutions it will make available guidelines for digitisation (quality, techniques, etc.) and storage (cf. Task 1.3) that fit into the existing existent institutional landscape and aim at complementing it. Second, 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 **(3.2.5) Geography and Demography**, quantitative approaches are stablished in the field and large amounts of digital geodata at different geographic scales are continuously produced and European countries use 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 **(3.2.6) Musicology** the TM will improve the digitisation and computer analysis of music scores (link) 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.

(3.2.7) Digital Humanities, a growing subject in Europe and the TM is an essential part of that growth, as an exemplary case of joining developments in computer science and CH. Connection with ongoing DH projects and institutions should be a priority for the TM.

Finally, **(3.2.8) Urban Studies** 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 & 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 – historical depth to present-day data collections. TM will enhance data collecting, processing and integration, for instance by enriching with unconsidered historical data collections. 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 and assist in the development of plans and strategies.

Targeted Achievement

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 realise the Pillar's overall objective. The urgency and difficulty of these challenges somewhat varies 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 Time Machine.

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 stablished, many of them in the form of RFC. Below the targeted achievements and milestones are listed and explained. 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 fit into existing practices. The RFC also needs to evaluate relevant technologies and recommend affordable technology that does not damage the objects and provide the best possible results. We aim to distribute cheap technology at 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 specialised 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 and then periodically. Aim of the review is to identify weaknesses in scanning recommendations, hardware, and software post-processing.

M1.1.3 - RFC on new scanning technologies. The cutting-edge technologies like automatic scanning machines with low human supervision, scanning robots and solutions for scanning films and books without the need to unroll/open them need to be considered and fostered by the TM. A specific scheme to incentivise these technologies will be created. We aim at an appropriate mix of dedicated specialised scanning centres and 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 in 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. Also, connection to long-term storage, e.g. DNA storage and selection of the most important data to be stored in such archives is an important design question.

M1.2.2- Implementation of TM Box: The data up to this point stored in different individual storage systems (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 priorities of objects to digitise**: Due to the massive amount of European CH and the different states of conservation, availability, and proprietary status, it makes sense to establish criteria to determine priorities of objects to be digitised. The rationale behind this hierarchy can be varied. It could make sense, for example, to offer priority to endangered objects that could be lost in the near future, but maybe accessibility and low difficulty in the digitisation process would provide more material faster and be a more efficient strategy. In any case, the different possibilities need to be carefully evaluated and a plan outlined.

M1.3.2- **RFC on models and formats**: Definition of guidelines and standards to follow 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 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 derivatives (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

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 quality human annotations, a proper interface is required. As many such interfaces currently exists, an assessment of the landscape is a prerequisite for the creation of a new TM interface

M2.1.2– **RFC on interface for annotation**: The created interface must allow for easy but complex annotation, that comply with the standards set for data modelling (M.1.3.2). The principles of human-computer interaction (taxonomy task 2.4) and previous user studies (M2.1.1) 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 so require, must be elaborated. The first step towards

this goal is an assessment of the current interfaces being used in the LTMs and other projects on digitisation of CH.

M2.2.2– RFC on user interface: The community will propose the features and requirements of the TM user interface. This milestone must be closely coordinated with the work on Pillar 3 "Exploitation Avenues", as the interface will be one of the main methods users will interact with the TM.

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. Considering the materials, the places of the LTM and Digitisation Hubs and the features of the languages a working plan of NLP tools development should be conceived.

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 M2.5.1 and 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 generalise 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, 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 adopted toward the need of large-scale machine learning. A likely addition is for example the option to provide gradient information of a specific module that is integrated using the API. This way also remote services can 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 such 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, also the generation and adaptation of existing models according to new evidence is in the focus of the 4D Simulator. Also, generation of virtual agent inhabiting the generated world is a focus of the 4D Simulator. An appropriate API 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. 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 enable intercommunication.

Main Areas Involved: 2.6 / 2.5 / 2.3

A2.6. Automatic Text Recognition

M2.6.1– RFC on Text Recognition (1): 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 (2): Following on 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 (1-2): 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, probable another will follow.

M2.7.2. RFC for music scores recognition (1-2): 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. Increased Acceptance of Quantitative Studies in Historical Research

M3.1.1. Call for quantitative historical research with TM Data Graph (1-3). The TM should incentivise 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 call 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 one 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 happen right after the publication of those tools, so that researchers can test and use them. The third call will happen in the final phase of the project.

M3.1.2. RFC TM tools for historical research. To engage researcher in social sciences and humanities are to productively use the Big Data of the Past, the TM can offer them a series of tools that facilitate the analysis. These tools will be enhanced by the Digital Content Processor and the Simulation Engines, which will enable scholar 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 models from outside the TM. A call to 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 outside models and tools. The simulation engines have the capacity to improve the performance and 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 that is 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 as the barriers, difficulties and success rate can only partially be assessed for a 10-year period and will depend on decisions by the TM operation structure.

	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029
A1.1 Digitisation Hubs	M.1.1.1. RFC Digitisation Hubs		M1.1.2 Implementation Digitisation Hubs				M1.1.3 RFC on new scanning technologies			
A1.2 TM Box		M1.2.1 RFC on TM Box		M1.2.2 Implementation of TM Box						
A1.3 TM Data Graph	M1.3.1 RFC on priorities									
	M1.3.2 RFC on models and formats									
A2.1 Annotation Interface	M2.1.1 User Studies	M2.1.2 RFC Annotation Platform	M2.1.3 Annotation Platform Implementation							
A2.2 User Interface	M2.2.1 User Studies		M2.2.2 RFC on User Platform		M2.2.3 Implementation					
A2.3 NLP Tools	M2.3.1 RFC Languages to address		M2.3.2 RFC NER	M2.3.2 RFC Orthographic Normalisation			M2.3.4 RFC Machine Translation			
A2.4 DCP	M2.4.1 RFC for DCP (1)		M2.4.1 RFC for DCP (2)				M2.4.1 RFC for DCP (3)			
A2.5 Simulation Engines	M2.5.1– RFC for TM APIs			M2.5.2 RFC LSIE		M2.5.3. RFC 4D Simulator	M2.5.4 RFC URE			
A2.6 Text Recognition	M2.6.1 RFC Text Recognition						M2.6.2 RFC Text Recognition (2)			
A2.7 Graphic Doc Recognition				M2.7.1 RFC Music recognition		M2.7.2 RFC map recognition				
A3.1Acceptance Quantitative Studies		M3.1.1 Call Research (1)			M3.1.2 RFC tools hist. research	M3.1.1 Call Research (2)			M3.1.1 Call Research (3)	
A3.2 Historical Simulation			M3.2.1 Call ABS				M3.2.2 RFC Simulation			
Funding		Yearly Open Calls								

Proposed Methodologies

Development of the state of the art and 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, the Requests for Comments (RFC). The structure and details on the RFC will be explained in the roadmap for Pillar 2 “Operation”. They are 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, call for papers and conferences will also enhance the communication and 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-proposal aim to attract bottom-up research proposals targeting specific milestones within a pre-specified time-frame. This research initiative can range from macro- to micro-level funding initiatives, from both national/regional/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 by Time Machine to a base project that was already 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 infrastructure and take advantage of the general Time Machine architecture when required. The periodical meetings (like the annual TM Conference) will be essential to share and evaluate the progress and the achievement of 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.

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 different 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 presented in our state of the art analysis 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). Yet, none of the funding sources is able to support such 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.

Using the proposed “on top funding” methodology, research labs, universities, and private companies are incentivised to integrate in the Time Machine Project. Doing so enables to combine developments already in progress by their existing funding scheme into 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 TM 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 is developing an active communication strategy to approach and receive 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 are involved in Pillar 1:

- 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 initiatives, programmes and projects on Cultural Heritage and Data Infrastructure for GLAM and the Humanities and the Time Machine LSRI needs to be inserted in this context in order to be part of a bigger and coherent European infrastructure. Collaboration with these institutions (for example, Europeana) are 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 capitalise 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 national still incipient initiatives at national level that try to push forward these principles (Germany, Portugal, etc.) but there is need to also 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 the 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 EU commission, TM repositories should be GDPR compliant. Respective measures for anonymisation of living persons must be set into place.

RISKS 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 planned 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 act as a hub to connect different institutions and projects will be able to work 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 contact with this field of research and foster innovative quantitative work on the Big Data of the Best with convincing results that make the case for the value of the TM for historical research clear. TM funding mechanisms incentivise cooperation with the TM.

(4) **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.