Research workflows, paradata, and information visualisation: feedback on an exploratory integration of issues and practices - MEMORIA IS

Iwona Dudek, Jean-Yves Blaise

Peer-reviewed & recommended by PCI Archaeo

Research workflows, paradata, and information visualisation: feedback on an exploratory integration of issues and practices - MEMORIA IS

Dudek Iwona*[1], Blaise Jean-Yves[1]

[1] UMR 3495 CNRS/MC MAP – Marseille, France,

*Corresponding author
Correspondence: iwona.dudek@map.cnrs.fr

ABSTRACT

The paper presents an exploratory web information system developed as a reaction to practical and epistemological questions, in the context of a scientific unit studying the architectural heritage (from both historical sciences perspective, and engineering sciences standpoint). The article presents the methodological and analytical potential of this system for the description, analysis and information sharing of research workflows.

The MEMORIA prototype is first and foremost an effort to build a tool that should help us to ensure the traceability, transmissibility and verifiability of scientific results and fulfil the challenges of open science (providing free access to the content produced). The specificity of the system is to empower a formal characterisation of processes that led to a particular research result by listing the most important elements necessary for a proper understanding of the result. An important point is the ambition to deploy visual interfaces providing access to resources and enabling direct analysis of the information collected. Ultimately, the project aims at depicting cognitive and methodological approaches behind scientific results using the possibilities offered by Information Visualisation.

The paper presents and defines the key concepts behind our approach and describes how they develop in practice. The theoretical aspects are illustrated with practical examples. The paper concludes with an analysis of the benefits and potential of the systematic approach to scientific process documentation that we introduce, highlighting its advantages and discussing its limitations.

Keywords: web-based information system, paradata, scientific results, research workflows, information visualisation, historical sciences

1. Introduction

Documentarists, archaeologists, cultural historians, computer scientists and ICT practitioners – more and more frequently stress the lack of reliable tools for documenting multiple aspects of research processes (i.e. its intellectual/cognitive, methodological, technological, tacit elements). The research initiative that we propose to discuss bases on the idea that beyond metadata describing outputs themselves, the scientific community is awaiting for 'paradata' that would include means to ensure traceability, verifiability and comparability of research workflows. This means sharing procedural knowledge and it is known not to be a straightforward task.

Doing things does not require the same skills as careful documenting, especially when the purpose of documentation is to picture comprehensively, the *how* and *why* of the overall 'production' process. In the *Journal of Documentation* Swedish researchers analyse how archaeologists document their work practices within archaeological reports [27]. They stress that sharing procedural knowledge is difficult "*… because of the lack of experience in thinking of their doings in terms of documentation …"* [27, p. 1109]. Describing and documenting, like perhaps all competences, require organisation, structure and practice. Not surprisingly, then, practitioners themselves find it difficult to describe what they do, as this refers to their personal knowledge (acquired by learning) and recollections – two interrelated but distinct aspects of memory.

First of all, ***knowledge has to be acquired***. If someone has not learnt how to document, he or she will probably struggle to do it well. Documenting needs to be taught.

Secondly, different processes involves ***different types of knowledge***. Carper [13] identified four fundamental *patterns of knowing* (in nursing): empirical, personal, aesthetic, and ethical knowing. Others, following Michael Polanyi, classify knowledge into *explicit knowledge* (knowing-that) and *tacit knowledge* (knowing-how) [18, 24]. Depending on the problem we focus on, knowledge can be furthermore categorized as *declarative* or *procedural* [8, 11], *contextual* [5, 9], *somatic* [5, 23, 25], *a posteriori* or *a priori* [43], and so on. This diversity reflects the complex nature of a knowledge to be transmitted. It encompasses at the same time facts, models, taxonomies, know-how, intuition, ethical components of a given practice, perception, mind-body action and reaction, the ability to identify one's own individual biases affecting the quality of work, etc.

In addition, human memory is unreliable, imprecise and subject to distortions. Our ability to remember detailed information about what we are doing and why we are doing it is limited - in quantity, quality and time [37]. Some experts highlight the inevitability of forgetting – "*… [forgetting], is not a failure of memory, but a function of it …*", "*… The brain is always trying to forget the information it's already learnt …*" [22. p.13]. It is therefore difficult for one to reconstruct a sequence of actions (activity types, decisions, their cause and effect relationship, their temporal sequence, instruments employed, people involved, etc.) within a longer process.

Finally, numerous frontline practitioners tend to understand their practice as the 'doing'. In her thesis, Rosmary Doyle [14] explores the relationships between practice ('doing') and documenting activity in social work. She points out that social work practitioners consider 'describing and documenting' practices as "*… secondary, bureaucratic concerns, with no material effect on the core processes and outcomes..."* [14, p. II] of their work practice. The statement that such a state of affairs is also observable in other areas of human activity – including heritage sciences - is not likely to stir up much controversy.

This paper presents the methodological and analytical potential of the MEMORIA web-based information system for the description, analysis and information sharing of research workflows. The system has been conceived and developed as a reaction to the practical and epistemological questions, in the context of a scientific unit studying the architectural heritage. The main purpose of the system is to identify various results obtained through scientific studies and to describe formally the processes that led to a specific result. Elements taken into account include: structure of a research process, methods and approach used in the investigation, tools and instruments, actors and their roles within the process, scientific context (e.g., organisational framework, objectives), sources, inputs *(*e.g., *outputs produced previously and described by a specific process)*, dates …

Before detailing how the concept of paradata develops in practice in the aforementioned information system, it is important to clarify a set of basic concepts related to it.

1.1 Metadata, paradata, provenance, process, workflow, pipeline - terminological confusion

The concepts of *metadata*, *paradata, provenance, process, workflow and pipeline* are increasingly used across disciplines. However, these terms lack sharp and commonly accepted definitions.

1.1.1 Metadata, paradata

Recent study on the use of the '*paradata*' concept in a corpus of scholarly texts from archaeology and cultural-heritage studies [41] points out a considerable disagreement about *what paradata is* and *what its purpose might be.* Some consider *paradata* as a subset or a type of metadata [10, 26] - calling them *provenance metadata* [3], or *reference metadata* [29]. Others perceive them clearly as information distinct from *metadata* [30], or assess concepts of *metadata*, *paradata* and *provenance* as partially overlapping [29, 36]. However, a common consensus can be found: paradata refers to information about productive process.

In the context of our work, in line with the definition given by Australian Research Data Commons [3], we view '*metadata*' as ***data/information about an object or resource*** that describes its characteristics such as content, quality, format, location and data administrative information.

In turn, '*paradata'* should be understood as ***data/information about a process by which an outcome was achieved*** including: methods (why and how the object or resource was 'produced'), intellectual processes (decisions, reasoning paths, …) circumstances under which results have been achieved (scientific context of the process, organisational framework,…), and the resources engaged (what tools, sources, inputs and approaches were involved in the successive stages of object or resource production, who was involved and when, …).

1.1.2 Process, workflow and pipeline

A similar cloud of semantic ambiguity surrounds the concepts of *process, workflow and pipeline.* Depending on the sphere of application - business, art or a specific scientific field - these terms may carry different conceptual meaning that can range from abstract concepts (e.g., *workflow* which refers to enterprise activities [7] to very specific, tangible elements (e.g., *workflow and pipeline* referring to scripts or codes) [31, 33].

Often used as synonyms, these terms can be associated with tools (e.g., management of document flows), diagrams (e.g., timeline or concept flow diagrams), modelling methods, sequences of real life actions, etc. However what seems to be the common denominator, is that they all refer to ***a sequence of tasks leading from a trigger to an outcome***.

In the context of our work we will use these terms in the following sense:

- a '*process'* will be understood as **a *series of actions or steps taken in order to achieve a particular end.*** In other words, it is the actual, real work as it is done ('*doing'*).
- a '*workflow'* will be considered as a ***view or representation of actual work*** – just as a map is a representation of a territory.
- a 'pipeline' will refer to ***a linear predefined sequence of steps or instructions***.

It is important to underline that a *workflow*, as opposed to a *process* or a *pipeline*, is most often a selection (deliberate or not) of actions through which a piece of work passed from initiation to completion.

The formal descriptions of workflows may vary from a purely verbal, linear, textual form [16, 32], to diagrammatic multipath representations [1, 12, 32, 34]. Visual representation of processes by diagrams is a well-established practice. They are widespread in business [7] and scientific research [21], but they are also used in law [34], construction management [12], health risk communication [2], chronic disease care [35] or education [32].

Workflows are commonly used to keep a *record of what has been done*, or to visualise a series of steps on the way towards the *project objectives*, as well as current problems and potential disruptions. Their visual character helps to understand patterns and relationships between activities, improves the accuracy of quantitative reasoning and, therefore, may improve problem solving and decision-making.

Thanks to their inherent property of displaying and ordering the relations between data over time and facilitating contextual and targeted reading of the data/information they present, workflow diagrammatic representations can convey protocols progression more easily than words alone. Tufte

claims that: "... *well-designed graphics are far more effective than words in showing observations ...*" [46, p. 87]. Yet, one should not assume that all graphics are more clear and understandable than a text. In this context, the role of the information 'recipient' must not be overlooked. Research surveys indicate that individuals' interpretation of the information graphics depends on their expertise (familiarity with the topic and prior knowledge) and instruction (graphic interpretation skills) [2, 20, 40].

1.2 Statement of need

In 2014, as a result of the lack of systematic documentation and archiving of research results within our team, as well as of the fluctuation of the people involved in research works, we noticed certain partial but significant information losses. We lost information about what was done, by whom, what the original objectives were and what limitations were encountered, which files contain what, how to open them, and so on. Data and information collected in the course of numerous projects and results of their analyses at various stages of progress were - and unfortunately still are - organised and described in an idiosyncratic manner. The central research problem of the MEMORIA project[1], launched to address the above difficulties, is to elaborate methods and technical means to depict scientific results, with indicators (metadata and paradata) that would allow for a better understanding of our research results as well as of the processes through which these results were obtained, and to make this information searchable and available for others.

In the following sections we present the methodological approach we have developed and adopted to structure, preserve and analyse *metadata* and *paradata*. First, we define the key concepts behind our approach, and present objectives that guided us. Theoretical aspects are illustrated with practical examples in the *Results* section. Finally, we present the benefits and the potential of the approach, highlighting its advantages and discussing its limits.

2. Methods

Our approach to the preservation and analysis of metadata and paradata is based on the following principles:

- the system should describe the ***results of research activities*** - this is where we have to decide what we consider to be a 'result', what types of results should be taken into account and what is the relevant information (metadata) we should record,
- for each result we should be empowered with means to depict the ***process*** that led to it – a process should structure the key elements necessary for a proper understanding and verification of the result by presenting the succession of actions and decisions in the form of a workflow diagram enabling *focus-plus-context* interpretation,
- the information collected should be accessible and queryable. It should serve as a basis for ***comparative analyses*** building on *InfoViz* practices.

2.1 Results of research activities (output, composition, publication)

Research activities may lead to highly heterogeneous results ranging from scientific outputs *per se* to dissemination products. Our choice was to distinguish three types of results that we call respectively *outputs*, *compositions* and *publications*. They are characterised by a series of descriptors (metadata) describing both the container (e.g, title, creator(s), format), and the content (e.g, one or more objects of study, temporal coverage).

The conceptual differences between output, publication and composition are reflected in the categories of metadata describing them (Fig 1).

[1] The MEMORIA project is a long-term project initiated in 2014 by members of the MAP-Gamsau team (UMR 3495 CNRS/MC MAP); *<http://memoria.gamsau.archi.fr/projet/objectives.php?lang=en>*

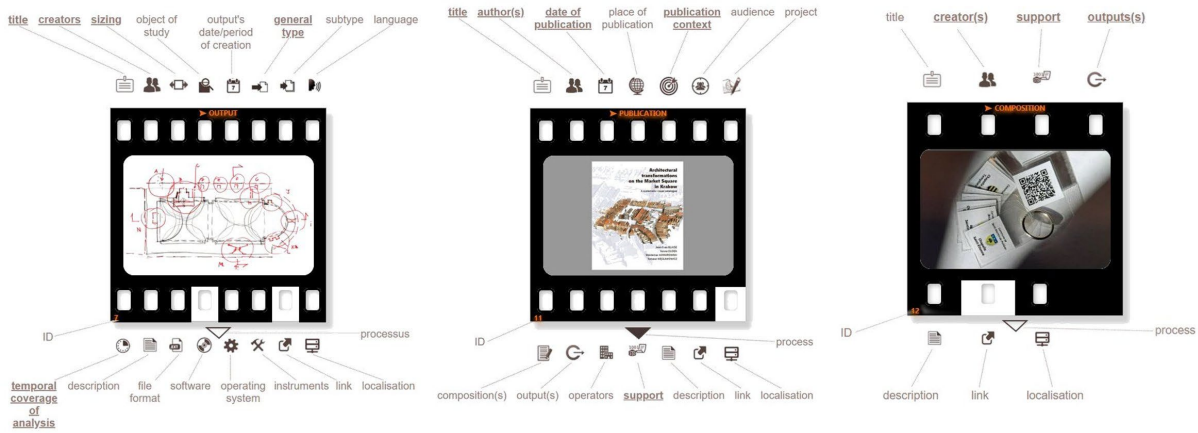


Figure 1 - The presence or absence of metadata for each element of the system is visually expressed by a 'film frame' metaphor which summarises the completion of the documentation effort.

Metadata provide only cursory information about what has been achieved, and where it is possibly available. Metadata alone do not allow to describe - and therefore to understand - what exactly the results are (e.g., data quality), how they should/can be interpreted, to which extent they can be exploited. In short, they do not allow an evaluation of the usefulness of the results nor their proper interpretation. This information is to be found in the *'process workflow'* (a formal description of a research protocol that led to an outcome's creation) and in the paradata that have been integrated into it (Fig. 6, 8).

2.2 Representing processes as workflows

In heritage sciences at large where observation, interpretation, subjective choices, specific and general knowledge, analogies, inference based on source selection intermingle throughout the scientific process, it is impossible to automatically trace the sequence of steps and decisions taken. Whether we choose to use an ethnic language or a diagrammatic representation to express the sequence of steps taken, the final result will show our *mental map* of the process.

Processes can be described by words. However, *"... a language [...] requires a linear progression of thoughts that may mask the interconnectedness of knowledge ..."* [6]. The examples of text-based workflows inherit the above-mentioned limitations - for example research workflows proposed by SSH Open Marketplace[2] [44] - designed to represent sequences of operations (pipelines) performed on research data during its lifecycle.

A research workflow is often a complex and non-linear array of sequences of steps. In that context diagrams allow the representation of relationships between objects, from a simple order to complex correlated relations (networks). They may be used to enhance *visual* (non-verbal) *reasoning skills* and exploit *iconic memory*. The desire to promote visual reasoning to support users' cognitive abilities was the starting point for the design of *MEMORIA workflow diagrams*.

2.3 Workflows from the InfoViz perspective

A process- *a series of actions or steps* – develops over time, therefore its visual representation (workflow), must be tailored to time-oriented data. In visual formalisms, time may be modelled in four different aspects: scale, scope, arrangement and viewpoint [1].

[2] The goal of SSH Open Marketplace scientific workflow is to enable sharing and re-use of workflows that are essentially a short, step-by-step guides. A workflow is composed of a sequence of paragraphs representing successive steps. Each step can be associated with the TaDIRAH [42] metadata and linked to various digital items (papers, web pages, reports) called *useful resources*. From the legibility point of view these workflows resemble a hypertext (with similar difficulties in grasping the whole problem at once).

The set of the most commonly used process-oriented diagrams includes: flowcharts, UML activity diagrams, lifecycle diagrams, timelines, Gantt charts, concept maps and mind maps (Fig 2). Some of them may be adapted to support different scales, scopes, arrangement or viewpoints, but each visual formalism has its own intrinsic limits.

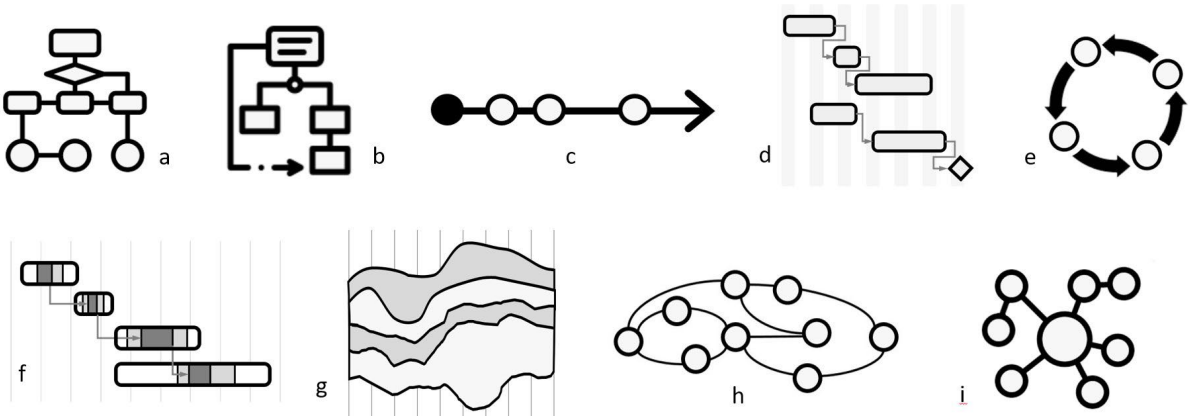


Figure 2 - Diagrams which are most frequently used to represent processes: ***a***) flowcharts, ***b***) UML activity diagrams, ***c***) timelines, ***d***) Gantt charts, ***e***) lifecycle diagrams ***f***) PlanningLines, ***g***) streamgraphs, ***h***) twist of pearls graph and ***i***) mind maps. [*cf.* 1, 34]

For example, a *flowchart* – the most commonly used process-oriented diagram - originally designed (in 1949) to plan computer algorithms, is tailored to represent decision making processes. It offers a terminology and 'syntax' dedicated to this purpose (e.g., input, process, decision, output, initialisation). Flowcharts and the family of similar process-oriented diagrammatic formalisms, including UML activity diagrams or DRACON diagrams, allow the representation of iterations, parallel and concurrent processing, but do not allow the distinction between ***repetitive*** or ***iterative*** activities (see Fig. 3). Since they were designed to plan computer algorithms they do not take into account actions that take place simultaneously or occur without any regular sequential order - typical of humans. Moreover, their graphical language may be off-putting. This may seem like an exaggeration, but graphics quality, is of major importance in helping to understand difficult schemes: "... *people may not accept or attend to graphics they dislike* ..." [2, p. 609]. Further analysis of the inherent limitations of the above visual formalisms is beyond the scope of this paper.

2.4 MEMORIA workflow diagrams

In order to avoid getting locked into pre-existing constraints, we decided to design *workflow diagrams* by thinking outside of the box. We have, however, built on our prior knowledge to draw on various existing visualisation formalisms. The main goal was to provide a diagrammatic representation that could serve as food for thought, enabling a *focus-plus-context*[3] interaction, preserving graphical simplicity and respecting the *data-ink ratio*[4] principle. The resulting diagram is a type of *cognitive map*, in other words: *a visual representation of a person's (or a group's) mental model for a given process* [19, online article]. Of all the formalisms known to us, concept maps are by far the ones closest to MEMORIA workflow diagrams. In order to highlight the specificity of MEMORIA workflow diagrams, we will describe them in comparison to this formalism.

Concept maps are visual formalisms for organising and representing knowledge in a form of a graph, where nodes represent concepts (names) and labelled and directed links - that organise and structure information - represent the relationships between them (linking words). Concepts are (in principle) represented in a downward-branching hierarchical structure – the most general concepts are arranged at

[3] The main idea of *focus plus context* visualizations is to provide viewers with the opportunity to see the object of interest presented in all its details (focus) while having the big picture - all the surrounding information (i.e. context) – available [28].

[4] A concept introduced by Edward Tufte. Data-ink is the "*non-erasable core of a graphic, the non-redundant ink arranged in response to variation in the numbers represented"*. If the data ink is removed from the image, the graphic loses its content. *Non-data-ink* does not transport the information, (e.g., background colour, edges ...). [45]

the top of the map and the more specific ones, are organised hierarchically below. Each node can have more than one parent, therefore nodes in a concept map are often cross-connected. The aim of concept maps is usually to explore relationships among several concepts [38].

The MEMORIA workflow diagram is basically a concept-map-based graph (DAG)[5], where nodes represent activities and the connections between them represent temporal relations. The key particularities of this graph may be summed up as follows:

- MEMORIA workflows diagrams are *time-oriented graphs* – conventionally, time expands in a rightward direction,
- links between nodes are not labelled, instead, their shape represents the type of time-oriented relationship (Fig 3),
- a workflow diagram is endowed with a set of symbols to annotate the process (e.g., *milestone notes* - a brief record of observations about a particular stage in a process, *infrastructure* used as a support during the process, link to the *preceding process(es)*,
- a formal identification of the moments in the workflow where results appear, differentiation of the type of these results (output, composition or publication),
- general data describing the process itself (e.g., name, project(s) within which the process is/was conducted, remarks) accessible behind icons.

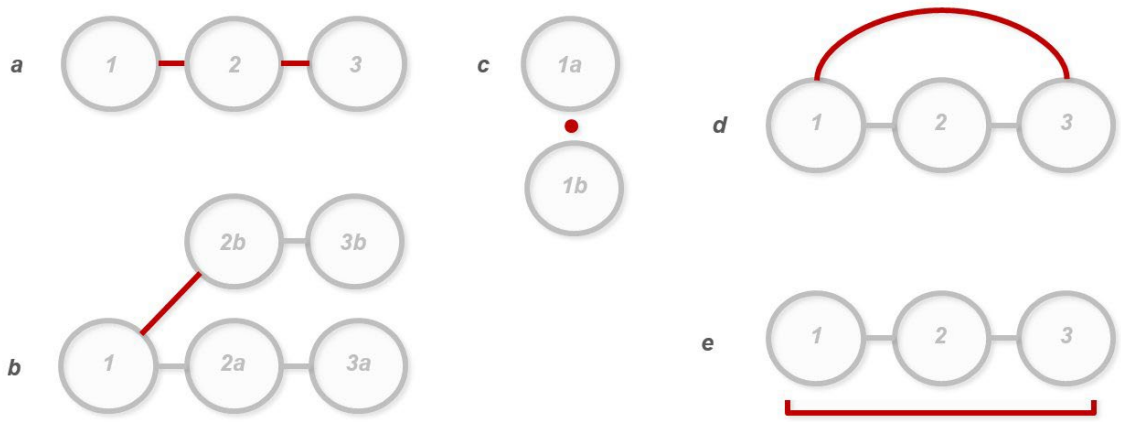


Figure 3 – Links between workflow nodes organise and structure time-oriented relations between activities, using following typology: **a)** *chain* - a sequence of a successive activities over time (displaying their order), **b)** *parallel sequences of activities* – activities that run concurrently in time, **c)** *knot of activities* - a set of activities that take place simultaneously or occur without any sequential order that could be determined, **d)** *iterative* sequences or activities (i.e. cyclic repetition of activities or their sequences on the same item, **e)** *repetitive* sequences or activities (i.e. repetition of the same operations on multiple items. In particular cases, when the information about the order of activities is missing, the order of execution of the activities within a process can remain unspecified.

Another noteworthy difference between MEMORIA workflow diagram and a conventional concept map lies in the visual encoding of the nodes.

2.4.1 Nodes (activities)

In a typical concept map a node is a labelled concept. By contrast, MEMORIA nodes encode the following information without the use of ethnic languages: activity type (glyph), its affiliation to one a particular group of activities (colour) and the presence of paradata provided for each individual activity (Fig 4).

[5] directed acyclic graph - a graph that does not have a closed loop

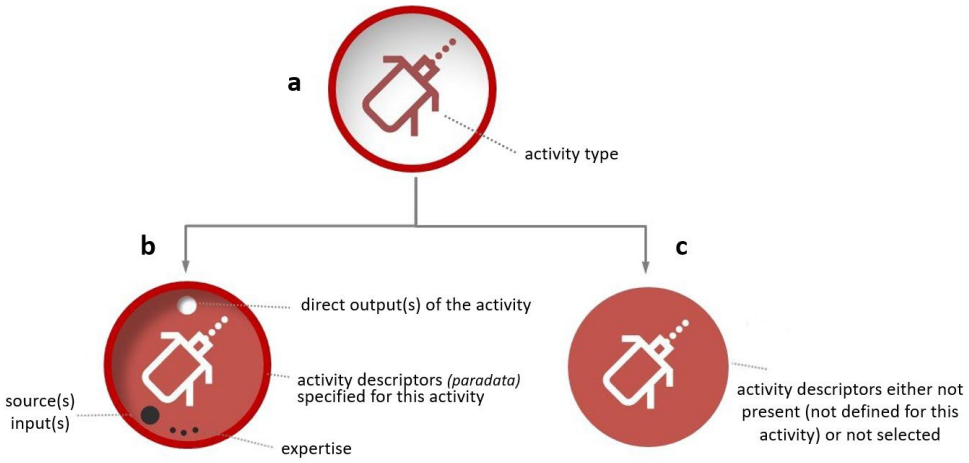


Figure 4 – An example of a MEMORIA workflow node (laser distance measurement activity) - red colour refers to the activity's group (data collection and acquisition). **a)** declared, but not yet described activity, **b-c)** described activities: the background colour refers to a given group of activity, the presence of visual elements indicates advancement in activity's characterisation (paradata).

Activities (nodes) are picked up by users via interactive '*wheel of activities*'- diagrams showing activities (with definitions and examples) distributed into five consistent groups (Fig 5). We have not based our model on any pre-existing example but on our data and our experience. All activities have been structured into five groups gathering goal-oriented activities dedicated to: A) data collection and acquisition, B) data filtering and treatement - transformation of the raw data into a form suitable for analysis, output production or finalisation needs, C) data analysis - encompassing methods of gaining of theoretical, explicit knowledge, D) added value procedural activities - the research phase centred on the use of procedural knowledge such as scientific procedures and technological protocols, E) finalisation - research process that is specifically undertaken in dissemination contexts, such as publication, communication, etc.

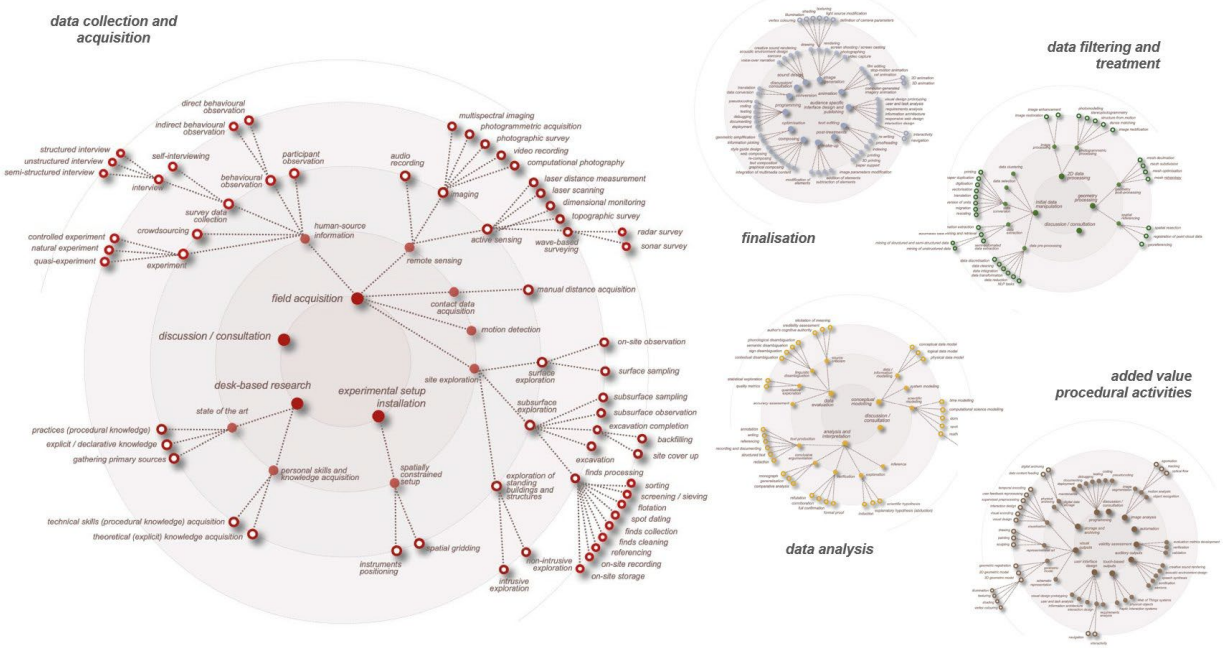


Figure 5 – Wheels of activities organise a set of activities specific to our laboratory in a systematic way that allows them to be easily retrieved and selected. Each activity is represented by a multidimensional icon indicating the category of the activity (colour), its type (glyph) and its hierarchical position. The result of this effort has been recently published: *<http://memoria.gamsau.archi.fr/projet/pdf/2022_MEMORIAact.pdf>*

Activities were collectively identified and structured during a long-term elicitation[6] cycle [15]. This elicitation step is a crucial and recurring stage in the MEMORIA IS (Information System) construction. It was recently undertaken in cooperation with a group of archaeologists and the system was extended to explore its potential in archaeology - more precisely in a phase of archaeological fieldworks. The division of activities into the above-mentioned five groups is sometimes perceived – especially in data science circles - as an approach based on *data lifecycle*. We comment on this aspect in *appendix 1.*

2.4.2 From the mental model of a process to its Memoria workflow diagram

Structuring one's own mental model of a process into a temporally organised diagram is a way for a researcher to interpret and question his/her own research protocol, but also to foster exchanges with other scientists on the basis of a shared description. However, as mentioned before, an individual's interpretation of the information graphics and their capability to model and structure their own knowledge depends on their graphic interpretation skills. Hopefully these skills can be learned, trained and refined.

At this stage of development, the system is functional and allows to describe and structure a process associated with the 'production' of one or more outcomes. Navigation through the *wheels* allows the selection of activities – all selected activities are stored within an 'activity tray', from where they can be dragged and dropped inside a composition grid and structured into a process by specifying relationships between activities (e.g., order of execution, iterative or repetitive segments, etc.) . From then on, activities dropped inside the composition grid can be documented with paradata (activity by activity). A workflow diagram is a sort of a backbone for the 'body of the process'. Paradata depicting each activity are anchored on it, they are freely accessible, searchable, editable (Fig 6). Organised by the means of a workflow, they are stored in an RDBMS (relational database management system). They can be queried, compared and visualised.

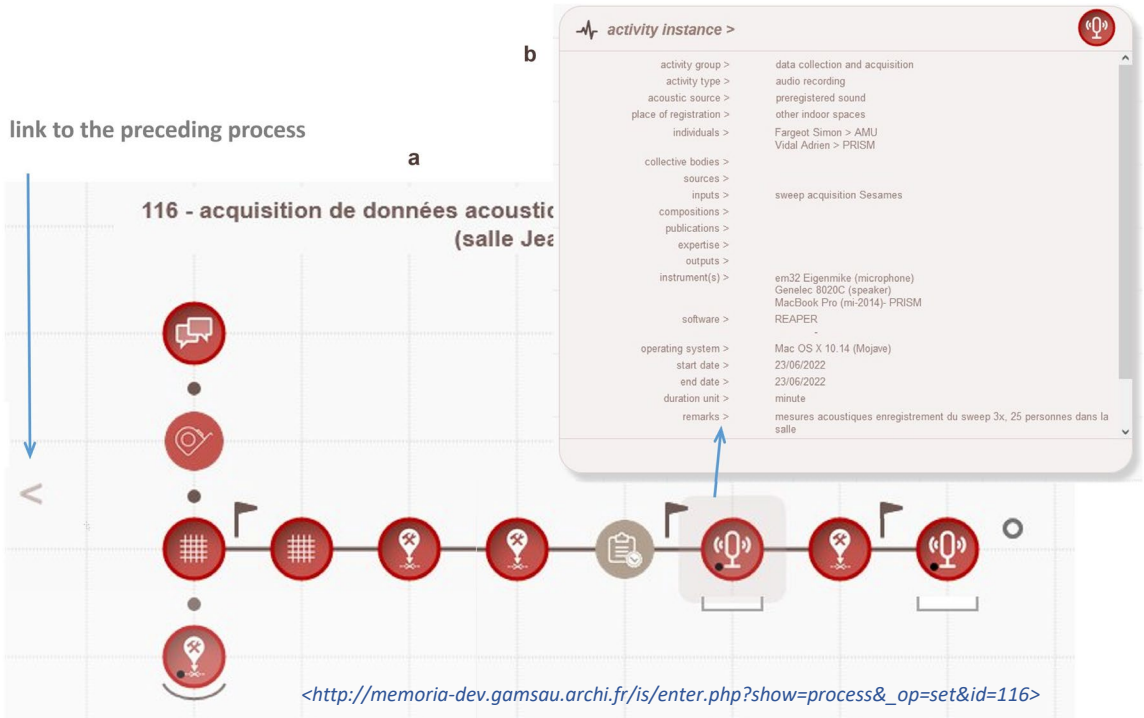


Figure 6 – The structure of the process acts as the structural backbone of the paradata declared and organised within each activity. **a)** Example of MEMORIA workflow diagram illustrating an acoustic data acquisition process. Instruments were positioned basing on a protocol applied a few weeks earlier. Upon casual inspection, this workflow showcases the linear structure of the process, composed predominantly of nodes corresponding to data acquisition. **b)** outputs' paradata declared within the highlighted activity (audio recording).

[6] The phase of elicitation was a long and complex process during which we sought to identify the structure of the professional practice of experts in the field. The process consisted of various phases in which forms of direct interaction with experts (e.g. focus groups), individual interviews or individual desk-based research were interwoven.

3. Results

The majority of the Memoria IS exploratory interface was developed and tested as part of the SESAMES ANR project[7]. During this period over 240 results - stemming from six independent projects - were identified and described (most of them are outputs). The mainstream of our work deals with study of architectural heritage, although not from a perspective of archaeology, so examples of workflows already present within the system may be thematically distant from archaeological practice – and therefore may be hard to comprehend. We have therefore chosen to overcome this complication by constructing a hypothetic project representing archaeological fieldwork based on a real-life example. We will use this example to concisely illustrate the particularity, usability and potential benefits of our exploratory system.

3.1 How to create a workflow diagram? From the identification of a result to workflow description.

The hypothetical example we will present below is based on the analysis of an actual archaeological fieldwork report[8].

3.1.1 Identification of results

On the basis of the report's content we have identified a set of ten outcomes (Fig 7). Nine of them, categorised as outputs and one as a publication, were described within the MEMORIA-sandbox platform – the IS familiarization tool. The next step, was to interpret the report in order to structure the excavation process into a *memoria* workflow diagram.

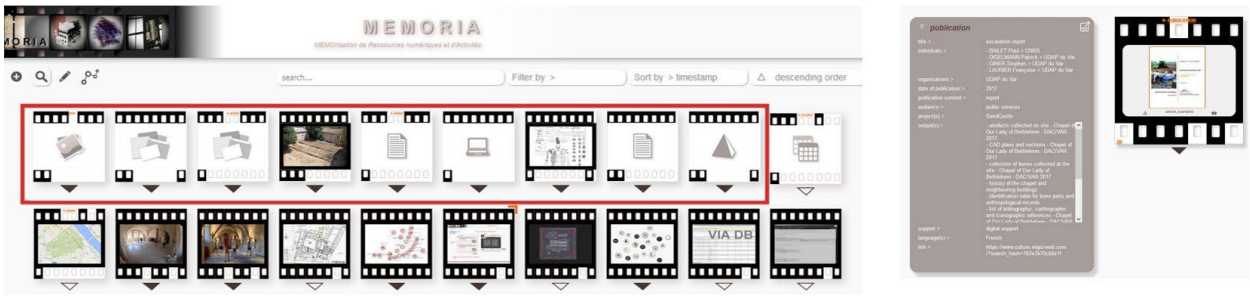


Figure 7 – Nine outputs identified (and partially described) for the illustrative example of archaeological fieldwork. On the right, metadata on the report, using the form of a *film metaphor*, illustrates the level of completion of the publication's description. *<https://sandbox.memoria.map.cnrs.fr/is/enter.php?show=publication&id=20>*

3.1.2 Workflow diagram as an answer to a question

Let us return again to the comparison of workflow with concept maps. Their inventor, J.D. Novak, points out that the sensible approach to construct concept maps is to work with reference to a *focus question* – "… a *particular question* we *seek to answer...*" [38]. A similar approach should be taken when trying to shape a workflow within the MEMORIA IS. Within our system, the focus question is: *how did I proceed to obtain this result? What were my subsequent steps?* In this hypothetical example the question became: *How did this group of archaeologists proceed to obtain these ten results? What were their subsequent steps?*

The answer to this question was particularly perplexing. The core of the report focuses on the presentation of the results of the operation and the research context (archaeological, geographical and geological ...). It contains exhaustive administrative data, but the section on methods and means is relatively short and contains not enough details to properly reconstruct the sequence and succession of the actions taken. The workflow diagram presented below is one of the numerous options we considered (Fig 8). This diagram certainly does not reflect the actual progress of the work, for which more detailed

[7] ANR (Agence Nationale de la Recherche) projet, ANR-18-CE38-0009-01 *<http://anr-sesames.map.cnrs.fr/index.html>*

[8] DIGELMAN, [39] The interpretation was done by I. Dudek (an architect with no practical experience in archaeological excavations), therefore the reader is asked to bear in mind the hypothetical and instructional nature of this demonstration, which is intended only to illustrate the functionality of the system.

information would be needed (i.e. field notes, layer forms ...). However, it is a version from which it is possible to start the reflection and discussion.

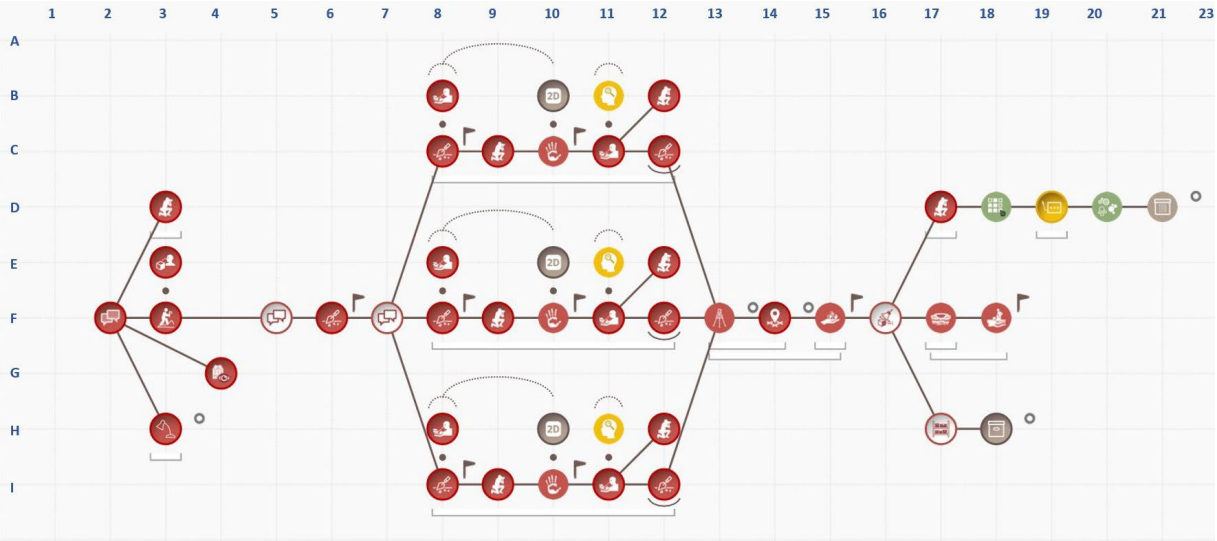


Figure 8 – MEMORIA workflow diagram - archaeological survey on the basis of an archaeological report <*https://sandbox.memoria.map.cnrs.fr/is/enter.php?show=process&_op=set&id=118*>. The grey rings indicate the moments at which the output data (identified in the system) appears. The story the diagram tells is detailed in *appendix 2*.

3.1.3 What might be the actual motives for describing the process?

An important decision is to make it clear for ourselves ***why*** and ***for whom*** we are documenting. What elements of the process we may consider as relevant (i.e. activities, milestone notes) and which paradata we will be willing to provide depends on the reason for documenting. The effort put into the description differ from case to case. If someone is documenting for themselves to prevent forgetting, then he/she will select the fragments that are important to him/her. The knowledge about a process will be structured differently for colleagues from the same discipline (who are familiar with the methods, objectives and approaches used) and for specialists from other disciplines (who know neither the objectives nor the specifics of the approach).

In the current state of implementation, the MEMORIA IS allows documentation of real-case (past and ongoing) processes. We identified two other potential uses for workflow diagrams - *'theoretical process'* and '*project process'*. The concept of *theoretical process* has a potential of passing on know-how and expertise of an individual or a research team - it could serve as a basis for tacit knowledge consolidation and transfer. One case study has already been tested[9].

Planned operations (*project process*) could also be recorded using a MEMORIA workflow diagram. A major opportunity in this case would be to enable documentation of the project during its execution using its project workflow diagram. A later comparison of the planning and execution diagrams could be used, for example, to highlight unanticipated elements. Both components – theoretical and project process - are pending implementation and testing.

3.1.4 How does the notion of metadata and paradata unfold in practical terms in the MEMORIA IS?

As outlined earlier, we consider *metadata* as data/information about characteristics of an object or resource. Their role in the system is to facilitate the findability and filtering of results, and to enable their comparative analysis. *Paradata* refer to the process of the production of results. Their function in the system is to allow the recording, sharing, explanation of the protocols used to produce the results, and to enable their comparative analysis.

The ultimate aim of these analyses would be enlarging epistemological awareness of the methods and approaches we use (identification of biases and errors for example), the influence of tools we use on the outputs we obtain, etc. In order to approach this goal, it is necessary to enter the sufficient amount of data into the system. However, the crucial and, possibly, inherent obstacle is that as people actively

[9] *https://sandbox.memoria.map.cnrs.fr/is/enter.php?show=process&_op=set&id=54*

involved in scientific research we often do not take time to document sufficiently our work. An in-depth analysis of this issue is beyond the scope of this document, but we will nevertheless shortly comment on it in the *Discussion* section.

3.1.5 What we have learnt from examining our case-study archaeological report?

Our attempt to describe this archaeological report in the form of a MEMORIA workflow diagram helped to highlight the nature and objectives of this document. It is a publication presenting several outputs to a given public (regional archaeological service) in the context of exploratory archaeological survey. It sums up results achieved in a systematic and organised way. An accurate and truthful description of the process based solely on the information contained in this type of document is impossible. The crux of the matter is that an archaeological report answers the questions posed *prior* to the operation – i.e. evaluation of the condition of the preserved archaeological layers and of the existence of previous occupation levels, supplementing the chronological phases already partially established. It therefore presents a selection of outcomes and summarises roughly the course of action.

The *metadata* describing the results of this operation is relatively complete (Fig 7). What is missing relates to file formats, instruments and software used to produce the outputs. The *paradata* that can be extracted from the report are more limited. Technological and methodological elements, such as the organisational and temporal pattern of the process, the role of each person in particular activities, the techniques used to uncover remains, the different phases of excavation strategy, the instruments and tools used, the obstacles encountered, the decisions taken, or the ethical and legal elements (legal restrictions, copyright) are not presented in a sufficiently precise manner in the report.

3.1.6 "To see is to forget the name of the thing one sees" [Paul Valéry] - the role of Information Visualisation within the system

The basic principle which has guided us from the very beginning of the project is based on the reinforcement and facilitation of non-verbal thinking in all phases of the system's use - in the phases associated with data feeding (filling in and recording the data), in the knowledge elicitation (process structuring), in the searching and filtering of the data and, above all, in the data analysis phase.

At the current stage of development, the MEMORIA IS helps the user to access structured metadata related to outputs and processes in read mode (the what, where, when and how of an output, etc.) and for all types of results to retrieve structured metadata in PDF format. The system provides data visualisation tools that can be used to exploit the content collected (metadata or paradata) to obtain a visual answer to a given question *such as: What are the main categories of activities that make up an individual's operational profile? What are the relative proportions of each group of activities inside a process, and how many inputs, outputs/results and sources are concerned?* (Fig 9) *What are the relations between the various processes conducted within a project?* (Fig 10),

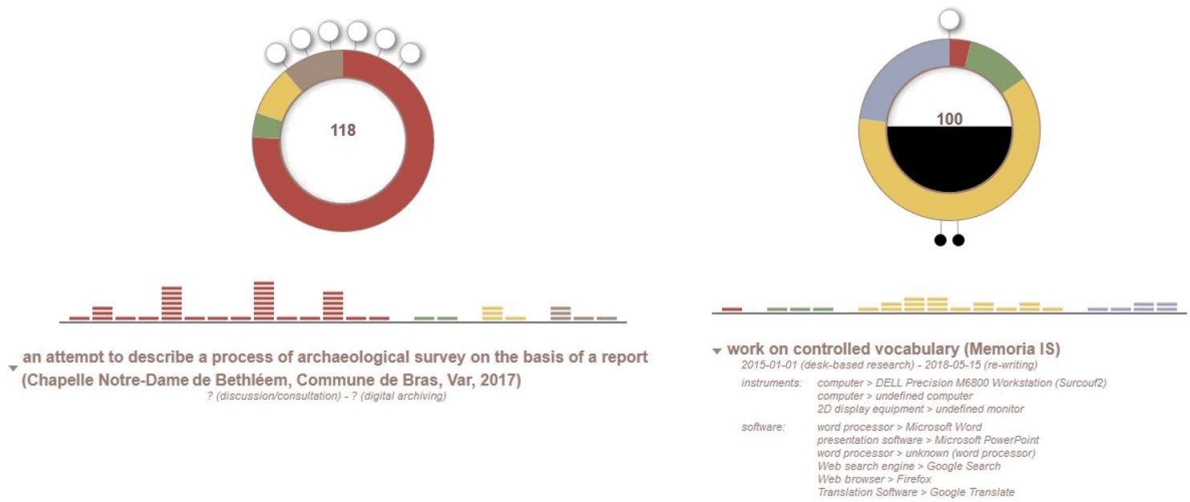


Figure 9 – Process as activities' proportion rings, one of the system's data visualisation tool designed to provide an answer to the question: *What are the relative proportions of each group of activities inside a process, and how many inputs, outputs/results and sources are concerned?* On the left, the case-study of archaeological fieldworks – six outputs identified

in this process, a prevailing number of data acquisition activities (red colour) with a predominance of photographic survey, excavation and subsurface observation activities, missing paradata concerning tools, dates, and instruments. *<https://sandbox.memoria.map.cnrs.fr/is/analyse.php?viz=1&v=118>*

On the right, the conceptual modelling process of the controlled vocabulary employed in the MEMORIA IS – the apparent predominance of data analysis (yellow colour) and huge quantity of sources used within the process (black feeling of the interior of the activities ring). *<https://sandbox.memoria.map.cnrs.fr/is/analyse.php?viz=1&v=100>*

Multiple interactions are possible with the activities' proportion rings, for example display of percentage of activities from various groups, links to and information about the process's results *(outputs, compositions, publications)* and about elements the process is based on (sources, inputs, expertise).

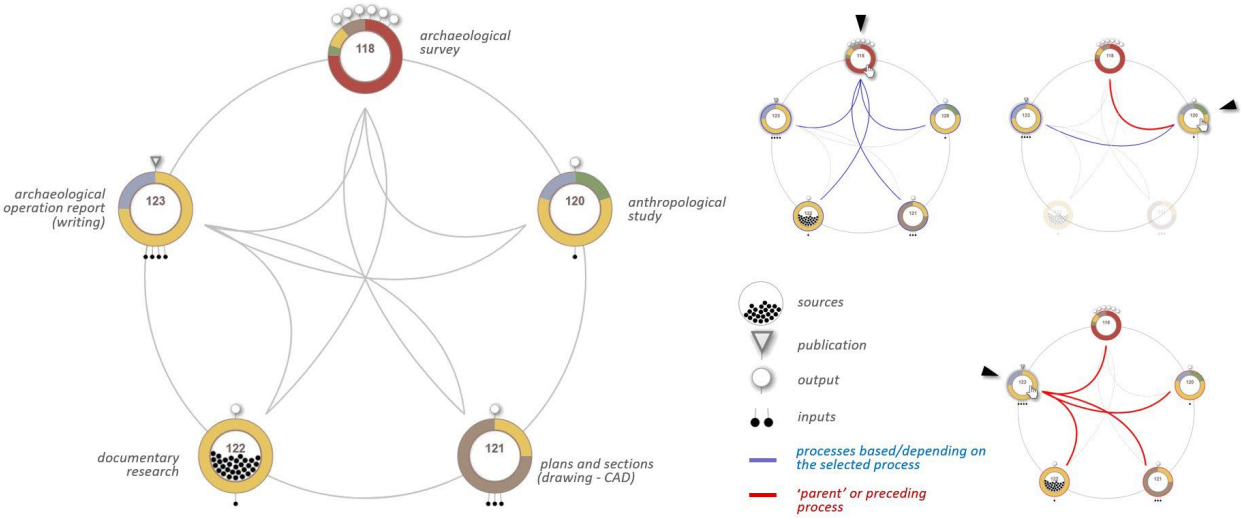


Figure 10 – Process variability chord diagram – a data visualisation tool displaying relations between processes conducted within a given project. On the left, the case-study of archaeological survey composed of five processes (four of them are non-data-driven study, drafted only as an example). Processes are represented as activities' proportion rings and ordered chronologically (clock wise) according to the start date of the first activity declared within a process.

On the right, examples of user interactions – an onmouseover highlights the processes declared as being related to the process that is being pointed to. In this case study, archaeological fieldworks produce outputs used as inputs in all subsequent processes. The anthropological study bases on archaeological fieldworks and serves the archaeological operation report. Report writing bases on all preceding processes.

<https://sandbox.memoria.map.cnrs.fr/is/analyse.php?viz=3&v=13>

The system interface is intended to provide the user with a visually stimulating environment, free of unnecessary embellishments (i.e. not carrying extra information), but showing variation in the data and drawing attention to the meaning and substance of the data.

A recent study on the documentation in archaeological field reports [27] points out that "… *reports have had a relatively poor renommée as being uninspiring to read [...] and according to critics, too often superficial, uninformative and of low scholarly value* ..." The scientific reports are probably not very thrilling either. We believe that a certain dose of visual stimulation can substantially deepen knowledge and understanding of processes and considerably reduce the discomfort barrier, making the work of the reporter and of the reader easier. This type of advantage, however, has its price - a detailed description of sequences of activities takes time. It requires determination and honesty, if the whole effort is to lead to understanding rather than to misinformation or camouflage of our failures.

4. Discussion

The MEMORIA exploratory prototype is, first and foremost, an effort to build a tool that should help us meet the requirements of scientific integrity (e.g., ensuring the traceability, transmissibility and verifiability of scientific results) and the challenges of open science (providing free access to the content produced), allowing methodological comparisons within the framework of intra-, inter- and

multidisciplinary work. A central point of the approach is the will to develop visual interfaces providing access to resources and empowering a direct analysis of the information gathered.

The MEMORIA visual interface gives access to *metadata*, *paradata* (structured within workflow diagrams) and dynamic visualisation tools. It allows the human analyst to interact with complex, abundant data about research results and the history of their production. But any tool - like a knife or an information system - can be used properly or inappropriately, for good and bad purposes. And so the MEMORIA IS has its potential - without recognizing it, we would not have committed ourselves for eight years to its slow, successive construction – and, unsurprisingly, it also has limitations. It is difficult to list them all in an organised way, because whether one sees them as advantages or not can be relative. Positive aspects of the system and its benefits include:

- *documentary potential* – an archive of traces of our professional activity, allowing us to store and share thematically structured collections of metadata and paradata, which can be translated to other formats (e.g., XML, JSON, text),
- potential role of the *theoretical/tacit knowledge-sharing environment* - possibility of sharing experiences and practical (tacit) knowledge with others (*cf.* 'theoretical process'),
- *innovative/creative visual analytical framework* – using collected data/information to analyse relations, trends and exceptions using dynamic visualisation tools,
- *support for the planning of scientific protocols* based on records of previous experiments (*cf.* 'project process'),
- possibly a groundbreaking approach to *meaningful documenting*[10] - using the system can allow us to look at our own work from a different perspective. It helps to be aware of *a whole* research process, of what has been done or abandoned, and of how it was achieved. It puts a spotlight on the need to take notes focusing on choices during the execution ('doing') phase.

However, each of the abovementioned points presents some methodological and practical limitations. Some of these are linked to our early choices. Its novel character puts a potential user on a steep initial learning curve, one that requires an intellectual involvement based on one's deontology. Structuring the MEMORIA workflow diagrams requires being ready to model one's own knowledge and this exercise may be time-consuming. Additionally, potential users should be open to learn meaningfully, and this requires an emotional and intellectual involvement to gain new insights through an autonomous discovery.

It should not be expected that this approach will suit everyone. A variety of reasons can be mentioned, such as:

- individuals may vary in their ability to describe the protocols in which they participated (e.g., due to the dilemmas of what is important to be included or difficulty in coherently conveying complex information),
- users may lack experience in conceptual modelling or graphic interpretation skills,
- users may lack willingness to explore ways of passing their experience to others,
- they may simply lack reliability or accuracy[11].

The first and second set of difficulties is not insurmountable - these skills can be learned, trained and refined. Others limitations are more fundamental. The very fact of using an information system is, in a sense, a limitation and a difficulty in itself, related to the issue of the *longevity and sustainability of web-based information systems*. The constant development and evolution of technology requires continuous mobilisation to ensure functionality of such systems (e.g., maintenance, development, user support and funding).

The documentary potential of the MEMORIA system is inherently restricted by its *purely digital nature entailing a certain fragility* of the stored data. For decades we have been cautioned about the

[10] In reference to the notions of *rote learning* and *meaningful learning* [38]

[11] The use of collected *meta-* and *paradata* to ***design new scientific processes and protocols***, despite its potentially encouraging nature, can also have its dark side - e.g., the 'copy and paste' syndrome. We know from experience that the misuse of this procedure, which speeds up the work to some extent, leads to an ill-considered, automatic use of one's own or third-party contributions.

necessity of digitalisation. Now we face an ever-increasing volume of digital documents occupying the disks of servers around the world. The risks involved include ensuring the long-term accessibility and readability of this data. Digital data are not less fragile than analogue ones, they are just different[12]. The numerical/analogue data archiving procedures considered in the 1990s[13] could be a good solution, for example, for archiving export files in PDF format.

The MEMORIA system may play the role of a *knowledge-sharing and knowledge-generating environment.* This requires reliable data within the system, and the correctness of the data entered depends entirely on its users. So far, the willingness to pursue a systematic description of our work - even of parts of it – is not prevalent within our community. The main – and relatable - argument is lack of time. Still, it should be clear that without a documentary effort and without making the documented data and associated scientific protocols openly available, our achievements may fail to survive and serve others, or enter the bloodstream of science, nourish and revitalise it.

Our actions depend predominantly on an acute awareness of their purpose. For what purpose and to whom should the data we collect be useful? Why and for what purpose should we document?

Furthermore, all the data, information and publications being produced today will not be stored indefinitely. Not only because of the increased costs incurred to store and keep them continuously available, but for purely pragmatic reasons: not all files deserve to be stored [4]. Determining how to *sort the wheat from the chaff* is a very difficult question, beyond our competence and the scope of this paper - we have not reached that point yet. For now, we are looking for ways to encourage 'harvesting'.

P. Fox recommends to: "... *Develop incentives to induce data providers to develop metadata and data products that will be usable by both narrow initial users of data and the wider community of interdisciplinary users reusing data for other purposes*. ..." [17]. Among the incentives, he indicates wider data use and data citations. Among the deterrents, he proposes contingent funding and contingent publication of papers.

4.1 Necessity of 'sticks and carrots'

It is only fair to admit that documenting is not particularly exciting. The majority of people we consulted on this issue seemed visibly crestfallen when they heard about the need for record-keeping. Within the MEMORIA IS, we have tried to incorporate several incentives such as a neat working environment to support the intellectual effort of description, involving visual thinking, design of visualisation tools that provide answer to user's questions, or offering converted metadata in a stable, human-readable format (PDF). Further improvements are possible. Introduction of the possibility of grouping the completion of certain metadata/paradata common to various activities to allow swifter form filling. The paradata structured within workflow diagrams could be accessible as a text and read aloud using a text to speech voice reader for visually impaired persons.

Nevertheless, there are also some 'carrots' over which we have no influence, like those recommended by P. Fox, or their variations, for example establishing regulations enabling for the publication of open data enriched with how-to (i.e. paradata) to become an act of publication of equal or increased value as compared to the publication of an academic research paper.

The 'sticks' are also necessary, but these are not at all within our reach (e.g., authorisation of research structures on condition that they have internal regulations requiring members to work systematically with open data + paradata, support research and applications aimed at improving cross- and inter-disciplinary discovery and reuse of data and results [17]).

Without a subtle and well thought-out combination of reward and correction, to challenge the inherent human inertia.

By designing and developing MEMORIA IS, we have tried to address the problem of a shortage of metadata and paradata in documentation practices at the local level (our lab). Our aim is not to convince anyone to abandon existing documentation practices, but rather to encourage an interdisciplinary reflection on the use, in one way or another, of theoretical principles or practical solutions proposed by the system.

[12] One solar storm in the near or distant future may erase them.

[13] digital to microfilm for example

Data, scripts, code, and supplementary information availability

Link to the MEMORIA IS website: http://memoria-dev.gamsau.archi.fr/is/enter.php

Link to the MEMORIA 'sandbox' -a "hands on" platform that allows potential users of the MEMORIA IS to try out the system and to explore the notions of output, composition, publication, processes and activities, or expertise: https://sandbox.memoria.map.cnrs.fr/is/enter.php

Link to the MEMORIA project website: http://memoria-dev.gamsau.archi.fr/projet/objectives.php?lang=en

The classification of the 285 activities present in the MEMORIA information system in May 2022. (PDF): http://memoria.gamsau.archi.fr/projet/pdf/2022_MEMORIAact.pdf

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Funding

This work was supported by Département de la Recherche, de l'Enseignement Supérieur et de la Technologie, Ministère de la Culture / Department of Research, Higher Education and Technology, French Ministry of Culture (2016-2020) and Agence Nationale de la Recherche [ANR-18-CE38-0009-01, project SESAMES, 2019-2023].

References

1. Aigner, W., Miksch, S., Schumann H. and Tominski, CH. (2011). Visualization of Time-Oriented Data. London: Springer
2. Ancker, J.S., Senathirajah, Y., Kukafka, R., Starren, J.B. (2006). Design features of graphs in health risk communication: a systematic review. J Am Med Inform Assoc, 13(6): pp. 608–619. DOI: 10.1197/jamia.M2115
3. Australian Research Data Commons. (2020). Data Provenance Metadata: Builds Trust, credibility and Reproducibility. *<https://ardc.edu.au/article/data-provenance-metadata-builds-trust-credibility-and-reproducibility>*
4. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature **533**, 452–454 *https://doi.org/10.1038/533452a*
5. Bakken, R. and Dobbs, J. (2016). The Relevance of Four Types of Knowledge for Leader Preparation in Radically Different Settings: Reflections on Data from a Case Study in Qatar and Teaching at a United States Military Academy. Creighton Journal of Interdisciplinary Leadership, 2(2), pp. 17-23. DOI: 10.17062/CJIL.v2i2.47
6. Bank, C.G., Daxberger, H. (2020). Concept Maps for Structuring Instruction and as a Potential Assessment Tool in a Large Introductory Science Course, Journal of College Science Teaching—July/August 2020, Volume 49, Issue 6. *<https://www.nsta.org/journal-college-science-teaching/journal-college-science-teaching-julyaugust-2020/concept-maps#JCST_65_B9>*
7. Barga, R., Gannon, D. (2007). Scientific versus Business Workflows. In: Taylor, IJ.et al. Workflows for e-Science. London: Springer, pp.9-16. *<https://doi.org/10.1007/978-1-84628-757-2_2>*
8. Berge, T.T., van Hezewijk, R. (1999). Procedural and Declarative Knowledge an Evolutionary Perspective. Theory & Psychology, 9(5): pp. 605-624. DOI: 10.1177/0959354399095002
9. Brezillon, P. and Pomerol J.-CH., (1999). Contextual Knowledge and Proceduralized Context. In: AAAI Workshop on Modeling Context in AI Applications, Orlando: AAAI Press, pp.16-20. ⟨hal-01574756⟩

10. Brusaporci, S. (2017). The Importance of Being Honest: Issues of Transparency in Digital Visualization of Architectural Heritage. In: Ippolito A. (ed.) Handbook of Research on Emerging Technologies for Architectural and Archaeological Heritage. Hershey, Pennsylvania: IGI Global, pp. 94-131.
11. Burgin, M. (2016). Theory of Knowledge: Structures and Processes. Kackensack, NJ: World Scientific <*https://fr.scribd.com/document/388187020/Theory-of-Knowledge-Structures-and-Processes-pdf*>
12. Carlsson, L. (2012). Visual Planning in Construction – a study of its use in construction projects, Master of Science Thesis, Department of Real Estate and Construction Management, Civil Engineering and Urban Management, Architectural Design and Construction Project Management, Thesis no. 146, Stockholm. <*http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A530578&dswid=-9932*>
13. Carper, BA. (1978). Fundamental patterns of knowing in nursing. Advances in Nursing Science, 1(1): pp. 13-24. DOI: 10.1097/00012272-197810000-00004
14. Doyle, R. (2009). Doing, describing and documenting: inscription and practice in social work, PhD thesis, University of St Andrews, St Andrews. <*http://hdl.handle.net/10023/766*>
15. Dudek, I., Blaise, J.Y. (2019). Enabling the comparability of research workflows: a case study, Proceedings - CAA 2019 (in press) <*https://shs.hal.science/halshs-02927631*>
16. European Open science Cloud, (2022). How to create a workflow in the SSH Open Marketplace?, <*https://marketplace.sshopencloud.eu/workflow/hmGpmv*>
17. Fox, P. (2011). Data Management Considerations for the Data Life Cycle, NRC STS Panel. <*https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwin3ZGh78z-AhXxVKQEHUdfCJMQFnoECAoQAQ&url=https%3A%2F%2Fwww.eudat.eu%2Fsites%2Fdefault%2Ffiles%2FPeterFox.pdf&usg=AOvVaw0DfhHVNd5l73XHUCrHGiTT*>
18. Gamble, JG. (2020). Tacit vs explicit knowledge as antecedents for organizational change. Journal of Organizational Change Management, 33(6): pp. 1123-1141. DOI: 10.1108/JOCM-04-2020-0121
19. Gibbons, S. (2019). Cognitive Maps, Mind Maps, and Concept Maps: Definitions, Nielsen Norman Group, online <*https://www.nngroup.com/articles/cognitive-mind-concept/>, consultation* 03 August 2023
20. Glazer, N. (2011). Challenges with graph interpretation: a review of the literature. Studies in Science Education, 47(2): pp: 183–210. DOI: 10.1080/03057267.2011.605307
21. Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, MR., Peters, K. and Schober, D. (2020). FAIR Computational Workflows. Data Intelligence. 2 (1-2): pp. 108–121. DOI: https://doi.org/10.1162/dint_a_00033
22. Gravitz, L. (2019). The forgotten part of memory, Nature, 571(7766): pp.12-14. doi: 10.1038/d41586-019-02211-5
23. Green, J. (2002). Somatic Knowledge: The Body as Content and Methodology in Dance Education. Journal of Dance Education, 2(4): pp. 114-118. DOI: 10.1080/15290824.2002.10387219
24. Halonen, R. and Laukkanen, E. (2008). Managing tacit and explicit knowledge in organisational teams. In: Pichappan, P. and Abracham, A. 2008. Proceedings of Third IEEE International Conference on Digital Information Management (ICDIM), London: pp. 292--297. <*https://www.academia.edu/27098504/Managing_tacit_and_explicit_knowledge_in_organisational_teams*>
25. Horst, T.L. (2008). The Body in Adult Education: Introducing a Somatic Learning Model. Adult Education Research Conference (St. Louis, MO), <*https://newprairiepress.org/aerc/2008/papers/28*>
26. Huvila, I. (2022). Improving the usefulness of research data with better paradata. Open Information Science, 6(1): pp. 28-48. DOI: 10.1515/opis-2022-0129
27. Huvila, I., Sköld, O. and Börjesson, L. (2021). Documenting information making in archaeological field reports. *Journal of Documentation*, 77(5): pp. 1107-1127. DOI: 10.1108/JD-11-2020-0188
28. InfoVis:Wiki, (2013). *Focus-plus-Context,* Information Visualization community platform <*https://infovis-wiki.net/wiki/Focus-plus-Context*>
29. Karr, A.F. (2010). Metadata and Paradata: Information Collection and Potential Initiatives, US National Institute of Statistical Sciences Expert Panel Report, November 2010, <*https://www.niss.org/research/metadata-and-paradata-information-collection-and-potential-initiatives*>

30. Kreuter, F. and Casas-Cordero, C. (2010). Paradata. RatSWD Working Papers 136, German Data Forum (RatSWD). January 2010. *<https://ideas.repec.org/p/rsw/rswwps/rswwps136.html>*
31. Leipzig, J. (2017). A review of bioinformatic pipeline frameworks, Briefings in Bioinformatics, 18(3): pp. 530–53. DOI: 10.1093/bib/bbw020
32. Mauri, M., Briones, Á., Gobbo, B. and Colombo G. (2020). Research protocol diagrams as didactic tools to act critically in dataset design processes. In: Proceedings of INTED2020 Conference, Valencia: IATED, pp. 9034-9043. ISBN: 978-84-09-17939-8
33. Mazzu, G. (2022). Bioinformatics Pipeline & Tips For Faster Iterations. *<https://www.weka.io/blog/bioinformatics-pipeline/>*
34. McLachlan S. and Webley, L.C. (2021). Visualisation of law and legal Process: An opportunity missed. Information Visualization, 20(1): pp. 192-204. DOI: 10.1177/14738716211012608
35. McLachlan, S., Kyrimi, E., Daley, B.,Dube, K., Marsden, M., Finer, S., Hitman, G. and Fenton, N.E. (2020). Incorporating Clinical Decisions into Standardised Caremaps (No. 2745). In: Proceedings of the IEEE International Conference on Health Informatics. Oldenburg: IEEE, pp. 1-2. DOI: 10.1109/ICHI48887. 2020.9374381
36. Moreau, L., Groth, P. and Dong Huynh, T. (2016). Provenance: An Introduction to PROV, W3C PROV Tutorial, 09 May 2016. *<https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwju07DOz6H7AhVEx4UKHWRaBJ0QFnoECAkQAQ&url=https%3A%2F%2Fwww.openphactsfoundation.org%2Fwp%2Fwp-content%2Fuploads%2F2016%2F05%2F140609_Cologne_IPAW-2014_Paul-Groth_Provenance.pdf&usg=AOvVaw2CMKLGQvFcuRhc44vnAVE2>*
37. Neisser, U. and Harsch, N. (1992). Phantom flashbulbs: False recollections of hearing the news about Challenger. In E. Winograd & U. Neisser, Affect and Accuracy in Recall: Studies of 'Flashbulb' Memories. Cambridge: Cambridge University Press, pp. 9-31. DOI: 10.1017/CBO9780511664069.003
38. Novak, JD. and J. Cañas, AS. (2008). The Theory Underlying Concept Maps and How to Construct and Use Them, Technical Report IHMC CmapTools *<https://cmap.ihmc.us/docs/theory-of-concept-maps>*
39. Digelman, P., (2017) Rapport d'opération archéologique, Bras, 12-21 Juin 2017. Patrimages No. RAP08651 *<http://patrimages.culture.gouv.fr/siteArcheo/16775>*
40. Phillips, R.J. (1997). Can Juniors Read Graphs? A Review and Analysis of Some Computer-Based Activities. *Journal of Information Technology for Teacher Education,* 6(1): pp. 49-58. https://doi.org/10.1080/14759399700200005
41. Sköld, O., Börjesson, L. and Huvila, I. (2022). Interrogating paradata. In: editor, Proceedings of CoLIS, the 11th International Conference on Conceptions of Library and Information Science, Information Research, 27(Special issue), paper colis2206. DOI: 10.47989/colis2206
42. TaDIRAH (Taxonomy of Digital Research Activities in the Humanities *<https://vocabs.dariah.eu/tadirah/en/>*
43. Tahko, T.E., (2011). A Priori and A Posteriori: A Bootstrapping Relationship, Metaphysica, 12(2): 151. DOI: 10.1007/s12133-011-0083-5
44. The SSH Open Marketplace, World Digital Library - Content Workflow, *<https://marketplace.sshopencloud.eu/training-material/vbob91>*
45. Tufte E.R. (1983). The Visual Display of Quantitative Data. Cheshire: Graphic Press LLC
46. Tufte E.R. (2006). The visual display of the quantitative information. Cheshire: Graphic Press LLC

Appendix

Appendix 1

groups of activities vs. data lifecycle

The division of activities into the five groups is sometimes perceived – especially in data science circles - as an approach based on *data lifecycle*[14]. However, our work was in no way inspired by this approach. Furthermore, a later examinantion of approaches in data lifecycle management allowed us to draw some interesting conclusions and to learn something new.

The practices focused on data lifecycles or persistence are conceptualised differently in different communities. Although they share similar objectives (i.e. *to* manage data throughout its existence), they vary in terms of the granularity of the problem presentation (containing a different number of stages), the type of data involved (e.g., business, government related data, scientific data) and the approaches (type of actions). Even a cursory analysis of a selection '*data lifecycle*' diagrams reveals the absence of consistent regularities of the succession of phases proposed in different contexts and for different data – only the position of a 'data acquisition steps' (red colour) might be seen as an example of a common trend. The presence of a *deletion* phase present only in some procedures is also revealing. In some data management groups the issue of data overflow and thus the need for regular data filtering and scheduled deletion is already noted and integrated, although only in some communities.

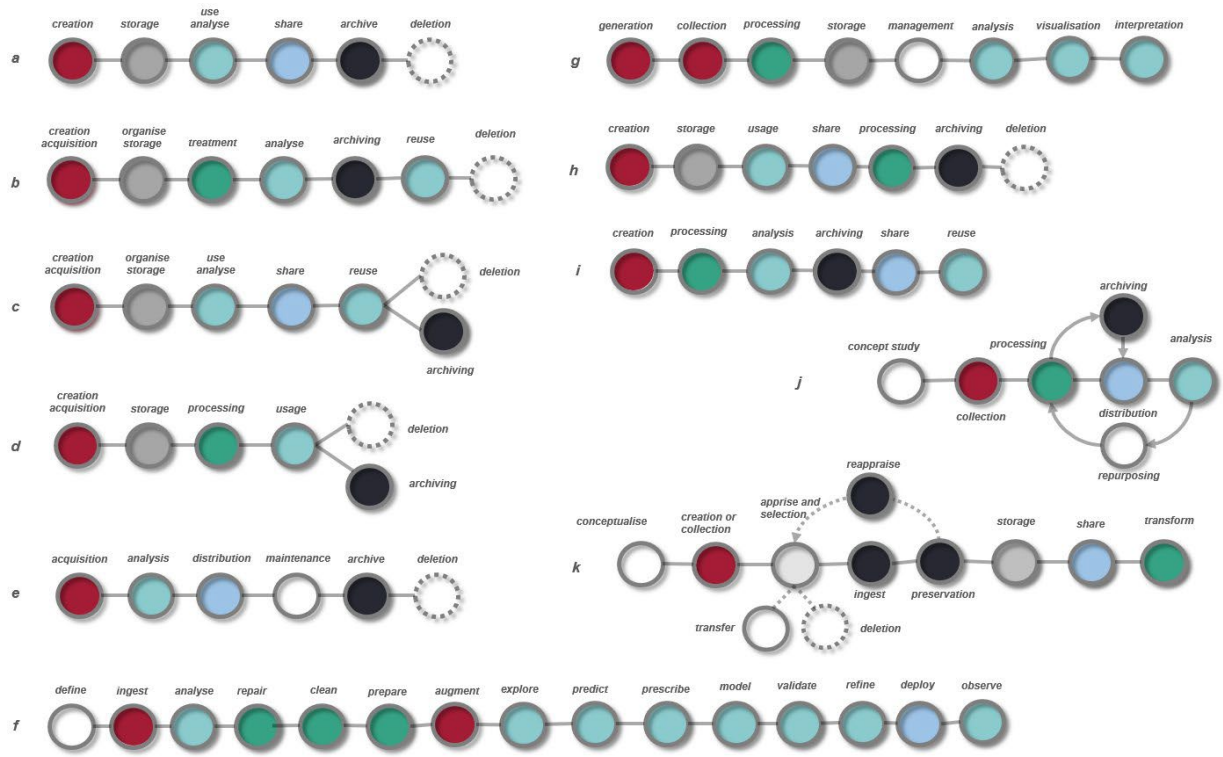
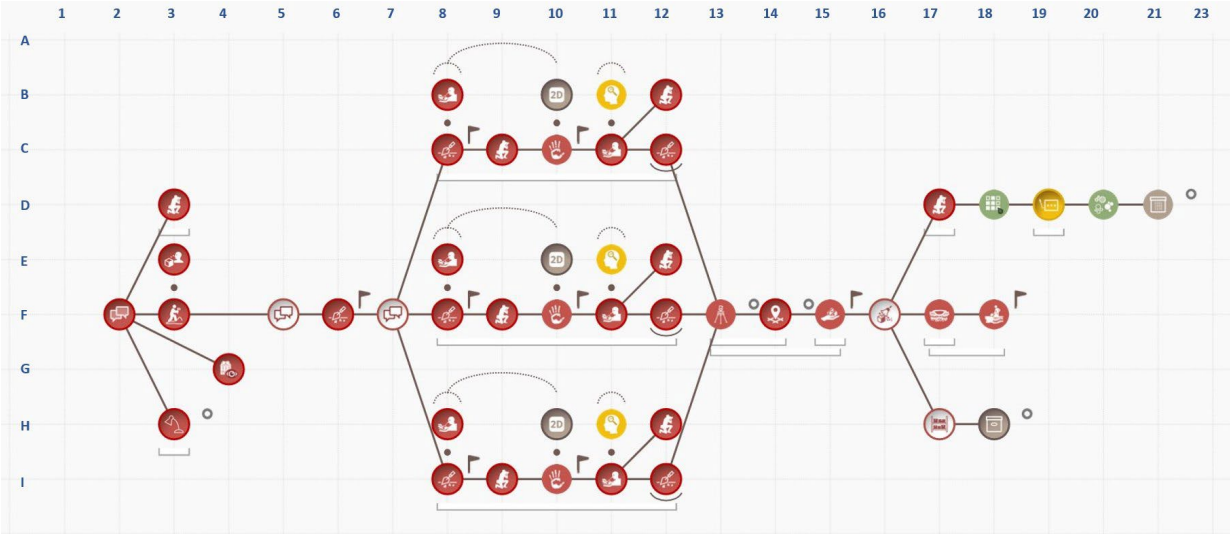


Figure 1a – Eleven graphs representing successive phases in data flow management (progression from left to right) put forward for: business data management (**a** - IBM, **b** - talented.com, **d** - startupgeek, **g**- Harvard Business School), geospatial data management (**e** - Sanborn), governmental data policy (**c** - NSW government), data management and analytics (**h** - ClicData, **f**- Trust Insights), research data (**i**- Nanyang Technical University), data curation (**k**- Digital Curation Centre), medical data (**j** - University of Nebraska). Original graphics have been redrawn and interpreted by the authors – colours were introduced to underline order of steps dedicated to similar goal-oriented activities (e.g., the 'ingest' phase in schemes k and f is denoted by a different colour, as the same term is used in these cases in other contexts. It is difficult to discern consistent regularities of the series of actions proposed in different contexts and for different data. The presence of a deletion phase present only in some procedures is also a tell-tale sign.

[14] Although this term is widely used, we will try to avoid it whenever possible, since it contains a semantically double-false element - *'lifecycle'*. In most cases, the term does not refer to any cyclical process, but only to a simple succession of actions/practices. Nor does it refer to 'life' as such, but rather to the *persistence* of data.

Appendix 2



Memoria workflow diagram - archaeological survey on the basis of an archaeological report <https://sandbox.memoria.map.cnrs.fr/is/enter.php?show=process&_op=set&id=118>. The grey circles indicate the moments at which the output data (identified in the system) appear. The diagram 'tells' the following story:

- **(F2)** the initial phase of the operation begins with a discussion between the participants (conjecture),
- from this point onwards, four parallel activities are undertaken: **(D3)** a photographic survey of the chapel, the facades of the houses in the hamlet, the site, ... - a repetitive activity before and during the excavation, **(F3)** surface exploration during the preliminary visit combined with **(E3)** potential on-site observation, **(G4)** non-intrusive exploration of the chapel - possibly before and during excavation, **(H3)** initiation of the documentary research - possibly before and during excavation,
- **(F5)** possible organisational discussions after the pre-visit, followed by **(F6)** outset of excavation – opening of three parallel trenches (mechanical shovel). From now on, excavation work is carried out in parallel (conjecture) in three trenches,
- work carried out in each of the trench concerned (hypothesis): **(C-F-I8)** iterative hand excavation combined with **(B-E-H8)** subsurface observation, followed by **(C-F-I9)** photographic documentation and **(C-F-I10/B-E-H10)** inventory of stratigraphic units (measurements and documentation). The whole sequence is repeated until an artefact is discovered. In this case finds are examined (first interpretations?) **(C-F-I11/B-E-H11)** photographed **(B-E-H12)**. Then follows a cautious hand excavation process **(C-F-I12),**
- in the case of finds collecting, the following activities are carried out (hypothesis): **(F13)** a topographic survey of the test pits, **(F14)** referencing localisation (of each artefact), **(F15)** and finds collection (repetitive sequences of activities) - the finds are then possibly cleaned **(F16)**,
- from this point, three parallel activities are undertaken (conjecture): **(F17)** site cover up and backfilling **(F18)** - carried out successively trench by trench (conjecture), **(D17)** photographic documentation of closing operations, followed by possible selection of photographs for storage **(D18)**, their annotation **(D19)**, classification of photographs taken during the operation **(D20)** and digital archiving of the entire collection **(D21)**, as well as on-site storage of finds **(H17)** and physical archiving **(H18)**.