Enabling the comparability of research workflows: a case study

Iwona Dudek, Jean-Yves Blaise

Enabling the comparability of research workflows: a case study

I. Dudek, J.Y. Blaise
Centre National de la Recherche Scientifique
Campus CNRS, 31 chemin Joseph Aiguier - Bât. US
CS 70071
13402 Marseille cedex 09
France
e-mail: iwona.dudek@map.cnrs.fr, jean-yves.blaise@map.cnrs.fr

*Corresponding author: **Iwona Dudek** (iwona.dudek@map.cnrs.fr)*

Keywords: knowledge extraction, visual reasoning, scientific protocols.

As a result of the massive introduction of computer-assisted research workflows in and around the analysis of heritage items, we are today witnessing a blooming of highly specialized, and sometimes obscure for outsiders, data processing chains. Operations conducted on this or that site include a vast range of digitization activities and an equally vast array of post-processing roadmaps.
The amount, diversity, and sometimes complexity of these operations are definitely a challenging aspect of the heritage science community's move towards "more" digital data acquisition and processing. It hinders that community's ability to identify and share, beyond results, methods and argumentation. In particular, it jeopardizes its capacity to preserve and explain research processes on the long term, and therefore to ensure their reproducibility (obviously a key methodological issue if processes should fall within a "scientific" approach).
This paper presents the MEMORIA research, aimed at experimenting a practical solution for the formalization and intersubjective description of heritage science research workflows.
The initiative bases on the idea that, beyond metadata describing outputs themselves, the scientific community concerned is awaiting for means to ensure their verifiability, reproducibility and comparability. The paper focuses on two aspects:

- A real-case experimentation on a series of investigations we have conducted on the historical centre of Krakow (Poland) over the last 15 years.
- A feedback on difficulties to foresee at a methodological level.

Introduction

Scientists and academics have over the last decades witnessed a profound shift in research practices induced by the growing influence of "digital technologies". As a result of the massive introduction of computer-assisted research workflows in and around the analysis of heritage items, we are today witnessing a blooming of highly specialised and sometimes obscure data processing chains.
Within our research unit, where constraints in relation with heritage sciences meet practices, protocols and tools stemming from the engineering and information sciences, we have already started to question our capacity to preserve and explain research processes on the long term (Doerr & LeBoeuf 2007) and – what is even more problematic – our capacity to ensure the verifiability and reproducibility of our own work (Rosnay & Musiani 2012; Guercio & Carloni 2015).

Such an observation, in the context of scientific activities addresses fundamental methodological and deontological questions. By "scientific activities" we mean here both activities which are carried out in the context of "hard" sciences and those which are more typical of the human sciences. Our approach to describing and recording scientific activities and processes is expected to be relevant both in science and scholarship. However, it is at this stage tested on the intersection of the disciplines and practices mentioned shortly earlier - Heritage Sciences, engineering, and information sciences. Therefore we shall make no claim that the approach has been validated on a larger scale.

One of the key issues that are discussed, is our capacity to perpetuate and transmit to future generations data as well as elements of knowledge. The project aims at the development of an experimental information system that would allow for the description and comparative analyses of our working methods, as well as help us gain a better understanding of their patterns of evolution. The mantra behind the MEMORIA[1] project is the idea that what needs to be formalised, shared, transmitted are not only results of research processes, but results along with their production processes.

In other words, the project aims at helping actors to preserve not only a digital resource, but also, and maybe above all, the way it was created: methods, protocols, choices and subjective interpretation layers that need to be worded if we wish to make of it a meaningful and reusable scientific document. In an application field where practices are often poorly documented, inferences sometimes subjective and instrumentations increasingly heterogeneous and pregnant, the problem is not trivial.

The contribution starts with a short "statement of need" and a brief definition of MEMORIA's research objectives. Next, we introduce the main concepts and principles of the experimental information system we develop. Then on a case study, we illustrate the methods and approaches we have adopted. In the last part, we briefly summarise some of the bottlenecks still ahead, and future works.

Statement of need

One of the basic deontological duties of a researcher apart from *knowledge development* is to share the results of his/her research work with the scientific community. This implies not only an effort to publish outcomes but also to ensure the intersubjectivity of workflows. Other researchers (and in particular future generations of scientists) should be given means to explore choices made during research processes so as to be empowered with means to re-question the data and the results.

As stated in the CNRS guide of integrity and responsibility in research practices: *The reliability of data produced by researchers relies on the adoption of appropriate protocols. Procedures to generate data must be described in clear and explicit terms so they can be replicated by other researchers and reused* (CNRS 2016).

Each scientific discipline has formed its own methods for documenting research protocols, but with the shift in research practices induced by the growing influence of digital technologies, we today notice significant changes in research pragmatics (*e.g.*, a renewal of workflows, a renewal of data exchange and publication paradigms and practices). If we want to keep control of what and how we do this or that investigation, we should try to understand those transformations (*e.g.*, the influence of new tools on the way we work, think, produce and share data) and support their comparative and cumulative analysis. This requires a completely new methodological

[1] The MEMORIA project is entirely conducted by our research unit with financial support from the Research Department of the French Ministry of Culture (DREST).

setting, grounded in an in-depth understanding and elicitation of the research workflows carried out in our domain.

Process workflows design is commonly used in economics. In this context a process is defined as: *combining various material inputs and immaterial inputs (plans, know-how) in order to make something for consumption (output)* (Kotler 2006). Hence, the objective of the production and analysis of workflow models – as defined by Kotler – is clearly profitability and gains in terms of productivity. This bias is worded by (Francis 2019) as follows: *A workflow model is the sequential series of tasks and decisions that make up a business process. Designing a workflow model lets business users see how a process works and helps them streamline and optimize it for best results and high efficiency.* A number of solutions aimed at the visualisation of workflow models have been developed in the InfoVis community (Aigner et al. 2011), but the service they offer is basically a visual monitoring of time-related variables (*e.g.*, outputs of an instrument, duration of a given activity).

The nature of outputs of such goal-compelled processes differs from results of research processes, where discovery, renewal of knowledge and intersubjectivity are crucial. The documentation of scientific workflows naturally requires to trace the provenance of inputs. But we also need to describe the experimental settings, as well as the methods and data sets used. Eventually, it is important to support scientists and scholars in their effort to analyse and visualise the pipelines they have followed (often trial-and-error, re-configurable ones, potentially open to serendipity). Modelling and sharing with the scientific community the above information is useful in a twofold manner: it is a way for a scientist to interpret and question his/her own results, but also to foster exchanges with other scientists on the basis of a robust, trustworthy and sharable description of the results.

The impact of computer-based pipelines on the modelling and visualisation of scientific workflows is coined by M. Atkinson (2017): W*ith the dramatic increase of primary data volumes and diversity in every domain, workflows play an ever more significant role, enabling researchers to formulate processing and analysis methods to extract latent information from multiple data sources and to exploit a very broad range of data and computational platforms.*

In other words, if there is a great need to document research protocols, it is even greater at a moment when scientists' practices and protocols include more and more data sets, tools and platforms that can possibly impact results.

A number of initiatives that address the issue on a general basis across scientific disciplines can be quoted. The *Research Object for Scholarly Communication (ROSC) Community Group Charter* introduced for instance in 2013 the notion of Research Objects, defined as "*research assets, including data used and generated in an investigation, methods used for producing the data, as well as people and organisations involved in the study*". The Researchobjects.org community site provides a list of initiatives related to that concept and associates the concept to FAIR principles. More examples can be quoted such as: *myExperiment* – a social web site for researchers sharing Research Objects, *VisTrails* – an open-source scientific workflow and provenance management system no longer maintained, however, *Discovery Net* – an e-Science platform based on a workflow model supporting the integration of distributed data sources and analytical tools.

One should however keep in mind that Heritage Sciences are (at least in part) *idiographic* disciplines (describing properties) and not *nomothetic* (establishing laws). Therefore the methods we implement do not overlap with those in 'hard' sciences (Bocheński 1968). The data at hand, with its load of uncertainties or a reasoning process that does not target one specific solution but accepts contradictions, are among the aspects to consider.

Hence, in the cultural area, specific challenges are raised – as exemplified in the development of the CIDOC CRM high-level ontology (Doerr 2003). The CIDOC CRM is a conceptual reference model providing an extensible ontology for concepts and information in cultural

heritage and museum documentation. Although there is an overlapping between that event-based reference model and what we are aiming at in the MEMORIA project, the overall aim of CIDOC-CRM and its extensions is to provide a reference model and information standard to describe data collections (CIDOC 2017). It puts the focus on semantic interoperability, whereas our concern at this stage is to picture human-based processes that may or may not bear connection with data as such.
Furthermore, one can observe that there is still a huge gap between the 'idea' that scientists should document their research protocols and their capacity to deal with the current jungle of formats, languages, schemas, specifications, data models, conceptual models, serializations, linked data end points, and so on.
Our position is that if we want to avoid the consequences of a second "dark age of digital archaeology" we must first contribute to a fine-grain analysis of our current research workflows, before jumping into description grids and this or that formalisation. We consider it is necessary to take the time it needs to be accurate on what we actually do, activity per activity, with an eye clearly open on the impact of computer-based settings. Then will come the time to invest in interoperability issues, on standards and reference models, with alignment strategies that we will only then be able to fully master.

Objectives of MEMORIA research

The general objective of the project is to depict a scientific result, especially in historical and heritage sciences, with indicators that would allow for a better understanding of the process through which the result was achieved. The main challenge is therefore to define methods and tools for the *elicitation*, *structuring*, *recording, and analysis* of research processes. What we await at the end of the day is to empower actors in science and scholarship to reflect on, and to cross-examine, their workflows. The approach we have adopted promotes the development of interactive graphical interfaces facilitating visual reasoning on our working methods, on techniques and tools we use, as well as on their evolution over time.

At this stage, we have completed the following steps:

- ***elicitation*** (*i.e.* knowledge elicitation, a phase of identification and structuring of knowledge not yet formalised, imperfectly formalised, or even ambiguous) (Shadbolt & Smart 2015; Brudge 1998; O'Hagan 2019) – the process through which activities have been differentiated and named, then described and exemplified (Dudek et al, 2015) (267 activities depicted to date).

- ***structuring*** and ***recording****:* a major "knowledge modelling and representation" effort has been made to filter the list of terms and concepts resulting from the elicitation step. This effort includes interpreting and evaluating the relationships between activities at the conceptual and practical levels. Potential relationships between activities were highlighted, and activities belonging to the same "concern" were grouped and categorised in an ontological structure identifying different levels of specialisation (Brewstera & O'Harab 2007; Peltoniemi 2008; Dudek & Blaise 2017). Activities have been distributed into 5 groups corresponding to different categories (data collection/acquisition, data filtering and treatment, data analysis, added value procedural activities, finalisation). Inside each group, activities are organised hierarchically from the general to the more specific (*e.g.*, field acquisition > remote sensing > imaging > photogrammetric acquisition). Each activity inside

the hierarchy, whatever its level of specialisation is, can be further documented thanks to a series of specific descriptors.
It is important to stress, that all details concerning actions mobilised within the research workflow are not intended to be registered. Ontologies we produce allow for a description of a 'process' with varying levels of granularity – conceived to fit the precision degree of the information one may dispose of - a certain level of abstraction and simplification of reality is necessary. Any model becomes a theoretical oxymoron – like a map without reduction of scale – if not bounded by a granularity, *i.e.* a complexity reduction bias (Joliveau 2007).

- ***analysis*** step: data visualisation may be defined as *a graphical display of abstract information for two purposes: sense-making and visual reasoning and communication* (Few 2014). We have designed a set of visual tools to convey abstract information privileging visual reasoning (*e.g.*, relations between activities within a '*process*' are represented as a chained structure). They provided support for communication during the collaborative knowledge elicitation step (*e.g.*, the explicit specification of the activities' ontological structure including concepts and the relationships between them, is represented in the form of a "wheel of activities", Fig. 1).They can now be reused in user interfaces.

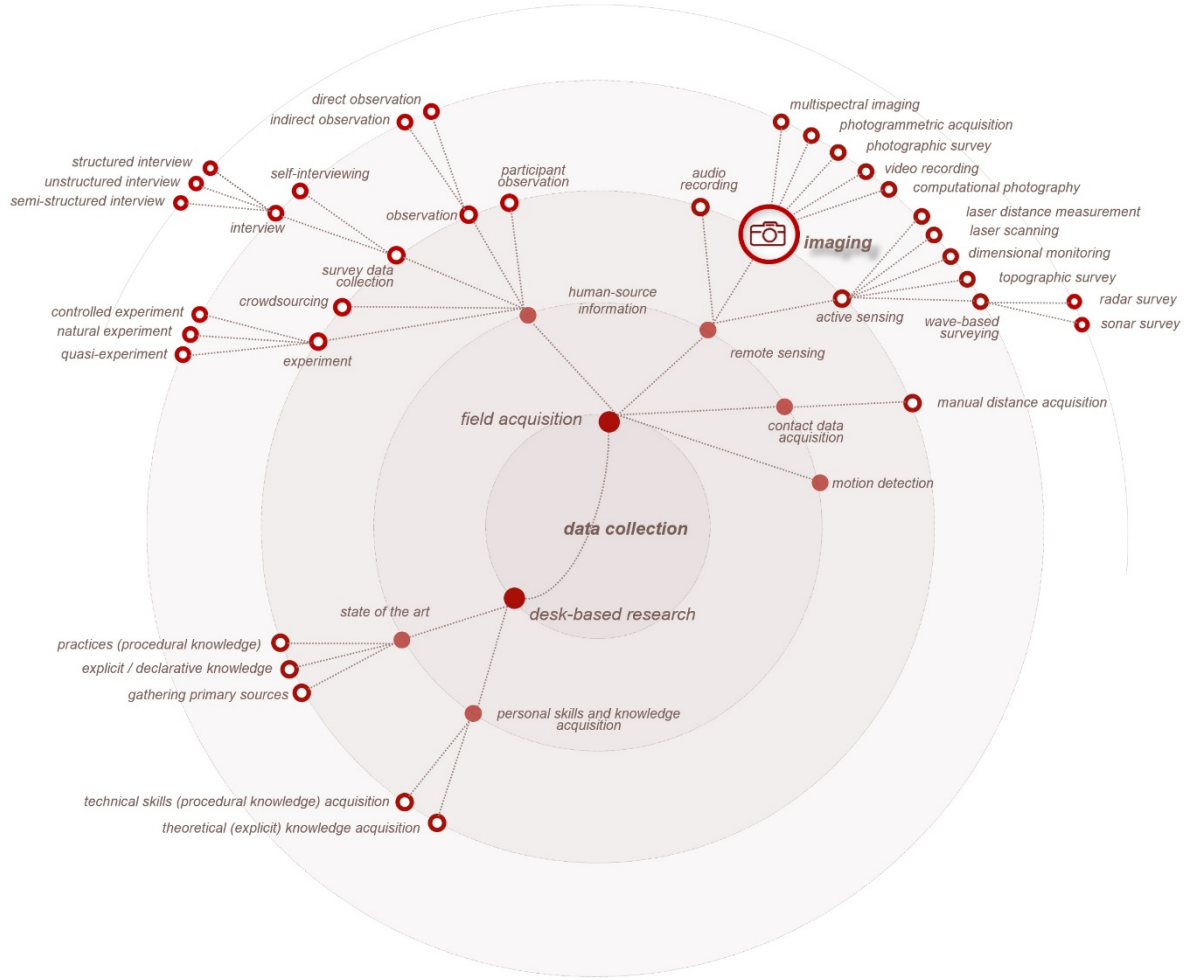


Fig. 1 The "wheel of activities" – a visualisation of the ontological structure of activities illustrated on the *data collection and acquisition* group of activities.

One of the crucial aspects of the MEMORIA project is the development of an interface building on a consistent visual language in order to improve the readability of queries and to facilitate the analysis of the collected data. What is meant here by *visual language*, is the result of a specific effort that interface designers can make to ensure consistency between semantics behind the information system and the modalities offered in terms of user interaction. For example, '*activities'* and '*processes'* are visualised by glyphs that reuse the colour codes and icons corresponding to the groups of activities. Each '*activity'* contributing to a '*process*' is represented by a multidimensional icon indicating the activity's category (colour) and its type (glyph) (Fig.1 and 5). Icons representing '*activities'* can then be grouped and structured into a '*process'* (Fig. 2).

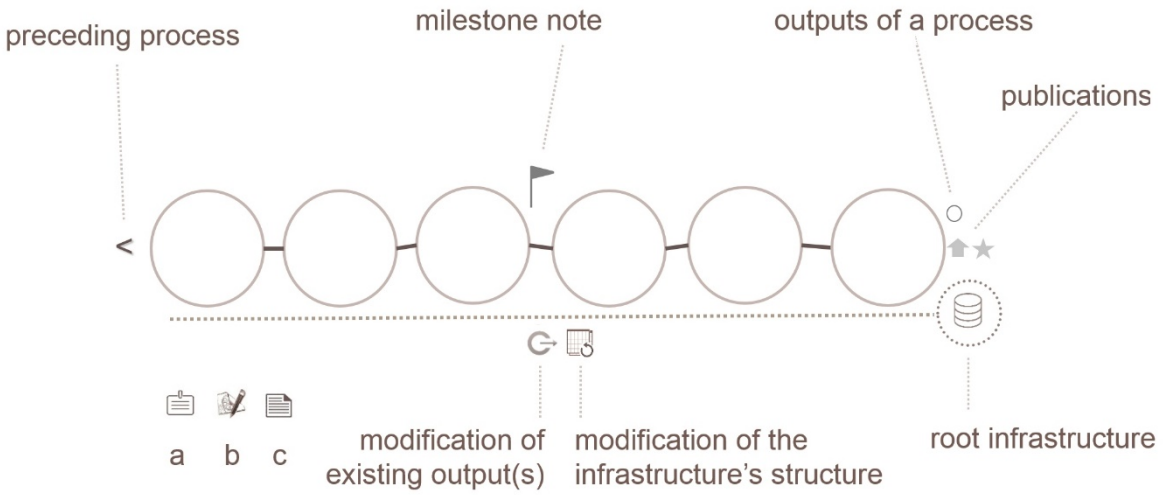


Fig. 2 A 'process' is a structured arrangement of successive activities conducted during a research workflow. It is identified under a unique name (a). It contains links to output(s) and publication(s) resulting from it, outputs modified during the process, infrastructures employed during the process pipeline, preceding processes, formal framework, *i.e.* a project (b) and free comments (c).

A set of symbols completes the graphical vocabulary associated with the representation of a '*process'* enhancing the visual identification of its particularities. The information on how the results were produced can therefore be visually summarised, and comparisons between distinct processes made.
We shall now exemplify this by detailing how the approach was applied to a case study.

The case study
Between 1997 and 2000 as a result of a study of the evolution of the old Town Hall of Krakow we created a collection of 3D models showing virtual reconstruction hypotheses corresponding to the most important development stages of this medieval architectural ensemble. What remains after this three-year experience, is a set of obsolete 3D Maya files, a collection of screenshots, as well as several articles and presentations.
Is it all that we can say about this output? Is it all that is left of this research work? Can't we describe the 'production' process of this output in a way that would allow for an interpretation

and a verification of the remaining results, and therefore an understanding of the cognitive and technical process behind the study?
Thanks to preliminary results of the MEMORIA project, we could begin defining, structuring and documenting the process through which we produced this output. As an introduction, we shall first clarify some fundamental principles and concepts of the MEMORIA IS.

MEMORIA IS – the basic concepts
As already mentioned, the information system we build aims at the description and comparative analysis of our working methods and of their evolution.
It is important to stress, that MEMORIA is not supposed to be a data storage system, a sort of new 'arch' preserving files containing the results of our work - for example, the Maya files containing 3D models produced during our study of the Krakow Town Halls' evolution. Its central role is to facilitate and expedite the identification, description and organisation of information on successive activities conducted during their production process. What we plan to store is the data on objectives we had in mind at the time we started that specific study, on methodological frameworks, on instruments, techniques and tools we used, etc. Once described, data and information are preserved (in ASCII format) and may be used to picture our ways of doing.
From this point of view, a resource resulting from a research activity that we call an '***output***', (might it be digital, digital-born, or totally analog) is the leading component of our system (Fig. 3). It can be a simple document (*e.g.*, a 3D model, a video) or a set of documents. Each *output* is inherently associated with one or several '***objects of study***' (*e.g.*, a building, a site, a movable object).

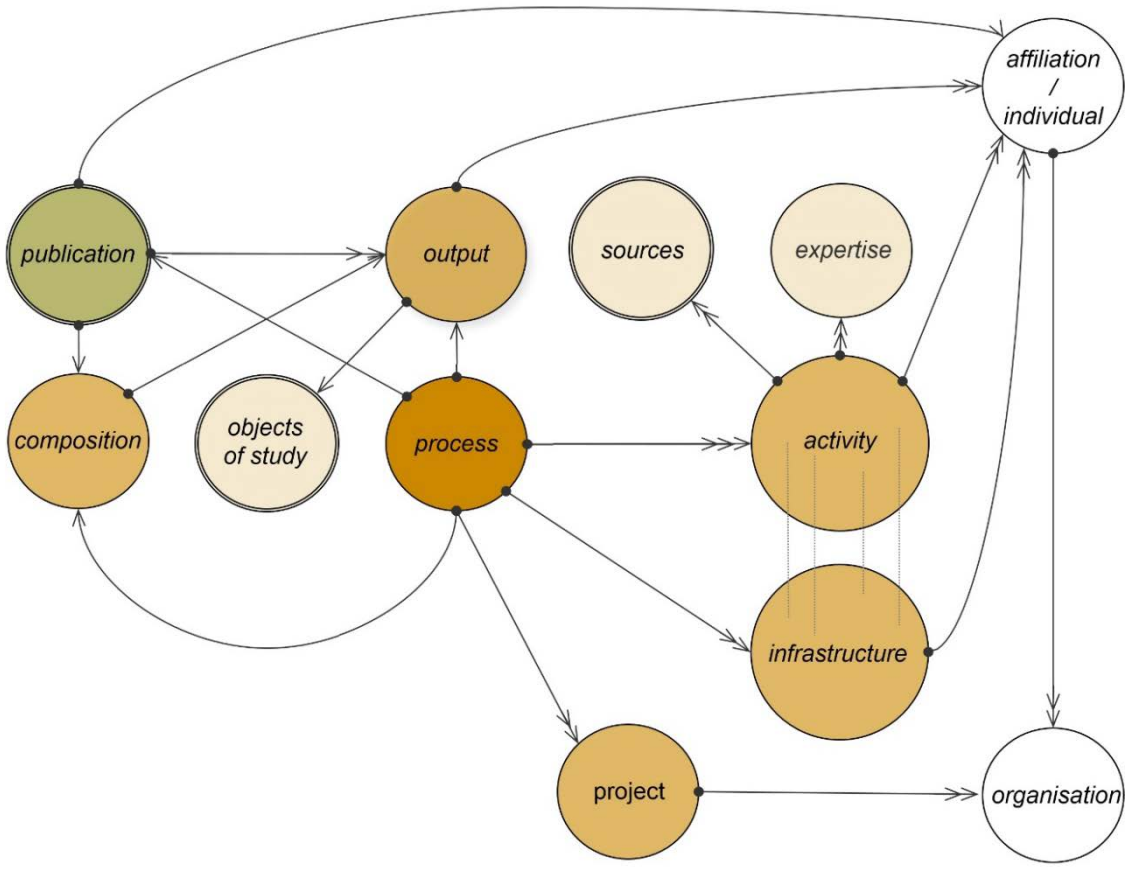


Fig. 3 Schema illustrating the basic concepts behind the MEMORIA IS and their relations.

Once produced, an '*output*' - and especially a digital output - does not become immutable. For various reasons, it may be subject to modifications (inside a new '***composition'***, *e.g.*, a poster, an open air-public presentation) or become a part of a '***publication***'. It seems meaningful for us to keep track of such repurposing processes.
Each 'output' is described as resulting from a '***process'*** (set of activities mobilised to produce an '*output')*. A 'process' may include one or several '*activities*' organised into a chained structure or left as an unarranged set of activities (*i.e.* a disordered ensemble) (Fig. 4).
An '***activity'*** encompasses a series of coherent actions undertaken to produce the '*output*' (*e.g.*, 3D modelling, data conversion, phonological disambiguation). It is the elementary component of any '*process*'. An '*activity*' can be based on one or more '***sources***' (*i.e.* external resources) or on previously produced '*outputs*'.
It has to be said that in some cases, the information one can pull together in order to depict a production process may be too scarce to allow for an exhaustive characterisation. There will be situations when we will lack information, as well as cases when some of the descriptors will be irrelevant. But lack of information *is* information *per se*. Such lacks are meaningful and may be exploited so as to visualise levels of incompleteness of the information, and ultimately to gain a better understanding of the nature of such shortfalls.

Configuring a '*process*'
The description of a '*process*' always starts with the identification of an '*output*' the researcher considers as deserving to be documented in terms of production process. The '*output*' can be a final result (*e.g.*, a virtual reconstruction) or it can be an intermediate result (*e.g.*, a cloud of 3D points resulting from a survey process, the translation of a text source).
We begin by identifying all the sequential activities. Then we can organise them as a '*process*'. If the activities' chronological sequence is known they can be structured as a chain. If that sequence is unknown or uncertain, it may be best to leave activities unarranged. Obviously, activities can be organised in more complex arrangements (Fig. 4).

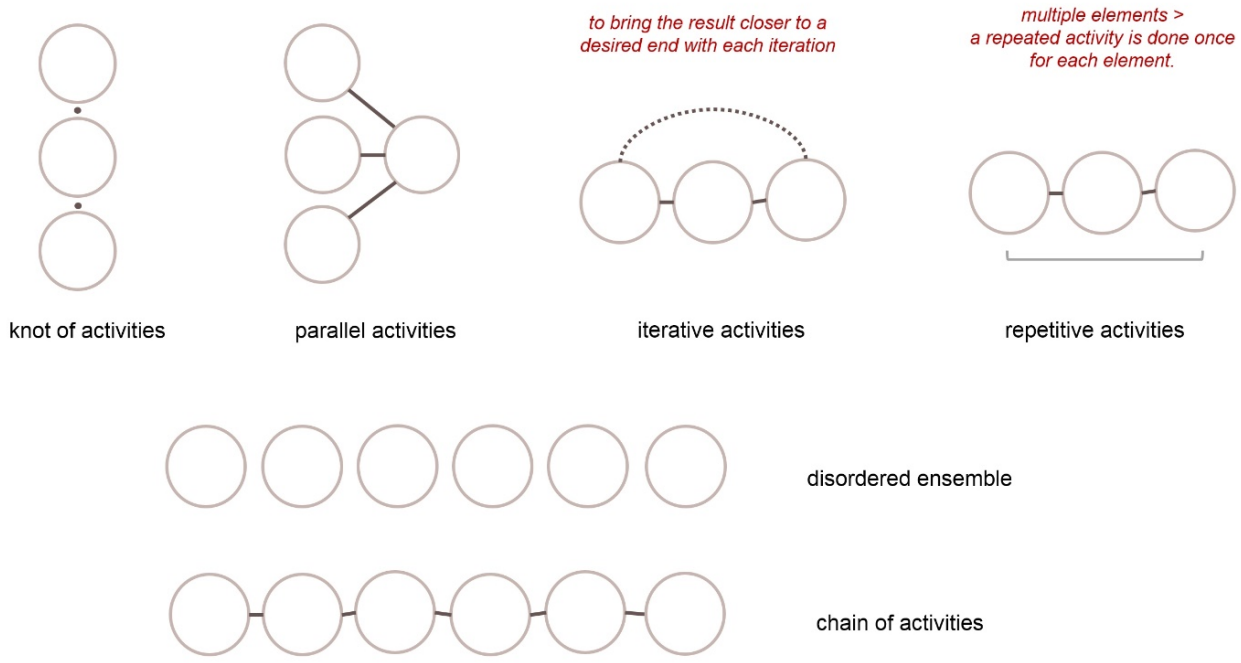


Fig. 4 An illustration of types of relations that can bind 'activities' inside a 'process'.

Regardless of the structure of a '*process*' additional information may be attached to it (*e.g.*, a link to one or several preceding processes, information about the root infrastructure employed during the process …).

When the structuration of a '*process*' is completed the description of individual activities can begin. Within a '*process*' each '*activity*' keeps track of (Fig. 5):

- institutional framework within which the work took place (creators, organisations)
- primary sources and experts' analyses,
- outputs, if resulting from the activity,
- techniques and tools used during the activity (instruments, software, ...),
- duration of the activity,
- recurrent or repetitive character of the activity.

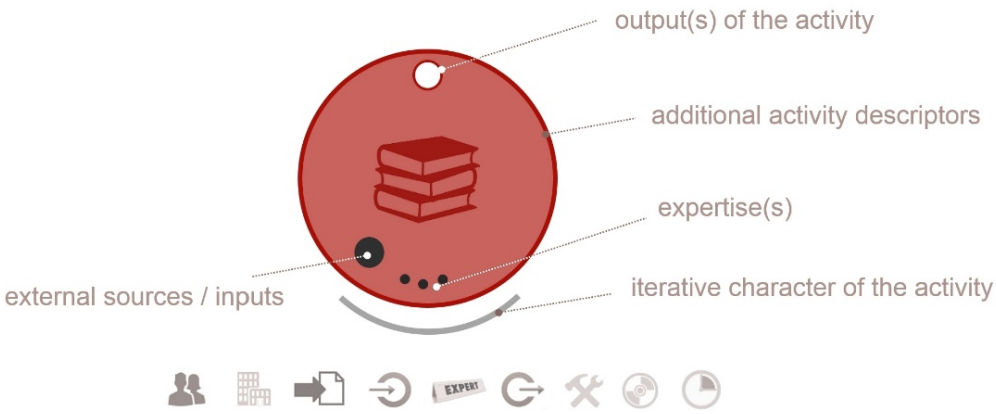


Fig. 5 The multidimensional icon used to map visually the descriptors of an 'activity'.

Visualisation step

The following graph (Fig. 6) sums up what we actually managed to remember about a research '*process*' that we completed 20 years ago – an analysis of the evolution of Krakow's old Town Hall.
The graph should be read as a text - from the left to the right (a click on an activity opens a window with a form containing a detailed description of the activity).
The whole process started with a data acquisition campaign (icons on the very left, red background). Primary and secondary sources collected during the initial step were photocopied or scanned. The following step was a repetitive phase during which information related to the Town Hall was manually extracted from each document and then analysed (elicitation of meaning, author's cognitive authority assessment, information and source credibility assessment) in an iterative cycle. The next knot of activities concerns data selection and

clustering (fifth column, from the left) – relevant pieces of data and information were filtered and distributed into five groups, each group corresponding to a given stage of evolution of the Town Hall.

The following step was dedicated to the translation of textual elements, and in parallel an activity was devoted to the acquisition of technical skills (procedural knowledge, namely training on a 3D modelling software). The subsequent activities corresponding to the 'data analysis' group involved domain knowledge modelling and inferences (identification and classification of architectural elementary entities corresponding to the morphology of the edifice, architectural analysis conducted by interpreting the sources gathered). The last activity before the 3D modelling stage was metrification (eight column, from the left). The closing, repetitive sequence of activities corresponds to the actual production of 3D models (last eleven columns). On the overall five distinct 3D models were produced.
The graph is quite explicit on the particularity of this final sequence of activities: a long and repetitive and/or iterative chain of technical activities, contrasting with previous segments.

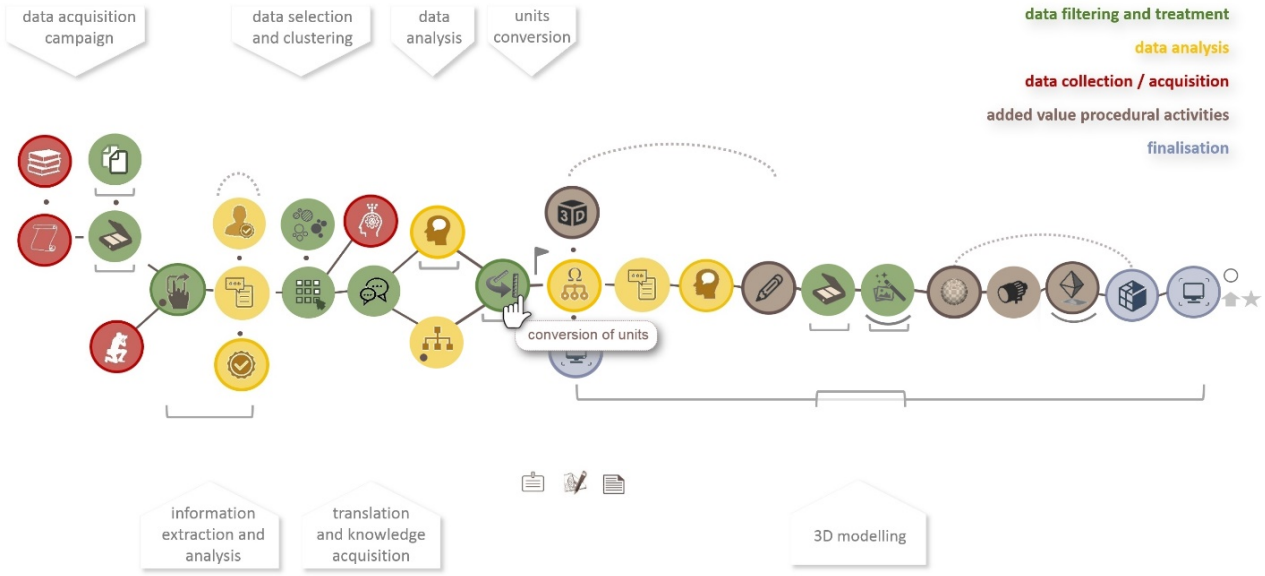


Fig. 6 Production process of 3D virtual reconstructions of the old Town Hall of Krakow: the *ordinal time* visualisation. Each circle corresponds to a given activity, and contains a glyph helping to identify it. The background colour associates it with one of the five 'groups of activities'. Lines below or above the glyphs are used to position recurrences: arches correspond to iterative activities or sequences of activities, horizontal "brackets" correspond to repetitive activities or sequences of activities.
An interactive version of figure 6 (PDF) can be downloaded using the following link
<http://www.map.cnrs.fr/BlackWhite/PubSc/CAA2019_processEX_fig6.pdf>

A process can be shown as a sequence of activities represented in *ordinal* time (only the order of the activities is shown, Fig. 6), or – if the information is available - in *ordered* time (time is composed of units, is continuous – the model of time behind classic calendars) (Fig. 7).

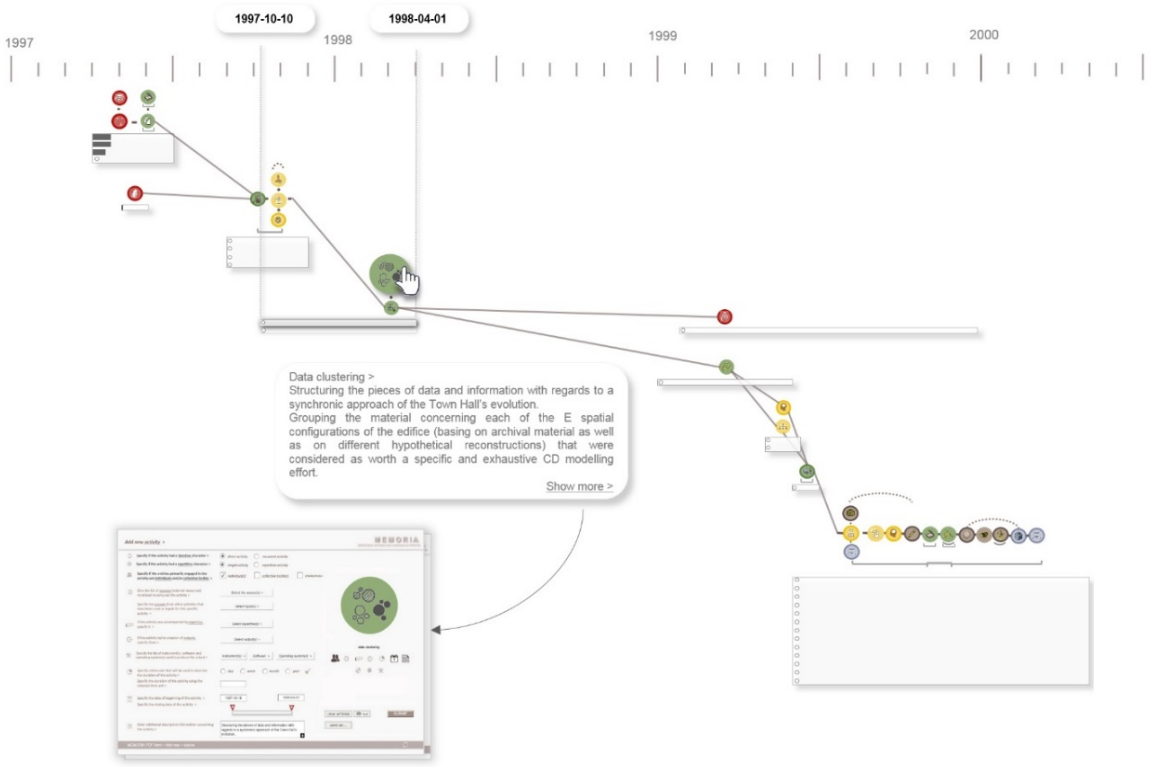


Fig. 7 Production process of 3D virtual reconstructions of the old Town Hall of Krakow: the *ordered time* visualisation.
The general estimations of the duration of activities are represented as temporal "bounding boxes". Here nine activities or sequences of activities are differentiated and ordered in time. Selecting a specific activity (here the 'data clustering' activity in temporal bounding box number 4) allows user to retrieve detailed information about the activity. Here in four cases the total duration of an individual activity is known (bounding boxes 1 and 2), however their precise anchoring inside the bounding boxes is not possible – lack of temporal anchor is represented as dark grey rectangles (duration interval symbol) positioned in the left part of the bounding boxes. When the duration of activities is not known a small grey circle is positioned in the left part of the bounding box.

Both visualisations give access to detailed data on each and every activity and ensure the navigation inside the information system. The difference between them is the presence of a temporal grid in the latter one. In this specific case study, we did lack a lot of detailed information about the dates and durations of this or that activity – as said before we were here experimenting the MEMORIA approach on an analysis conducted 20 years ago. In the majority of cases, we were able to provide only estimations of temporal "bounding boxes". One of the indication the *ordered* time visualisation underlines is the long lapse of time between the beginning and the end of the process (a consequence of the organisational framework).

Analysis phase

One of the main objectives of the MEMORIA approach is to enable *visual reasoning*. This type of reasoning affects nonverbal skills (*i.e.* visuospatial analysis, attention and non-verbal memory). It relies on perception rather than cognition. In this sense, our approach and in particular visualisation efforts intends to back up the analyst's cognitive capacity.
Once activities are structured into a '*process*' and relevant information about activities is provided, an analysis of processes is made possible. The analysis may be limited to an

individual process (individual analysis) or performed on a group of processes (comparative analysis).

Visual analysis of an individual process

Individual analysis of a '*process*' typically involves pattern search (*e.g.*, correspondence with established protocols, presence or absence of a given group of activities, search for a predominant group of activities inside a '*process*'[2].

For one feature that we wish to analyse several visual solutions are developed. The example below shows that recurrence patterns inside a process can be observed from different points of view, and lead to different findings (Fig. 8, 9).

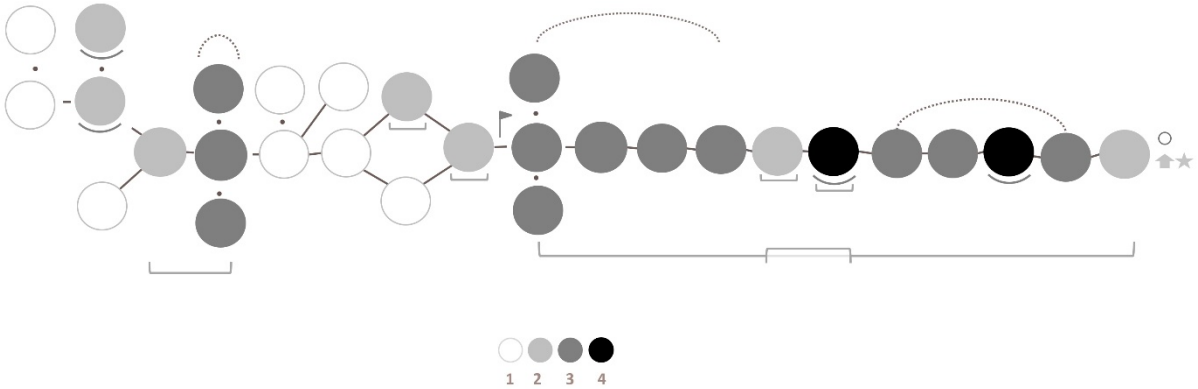
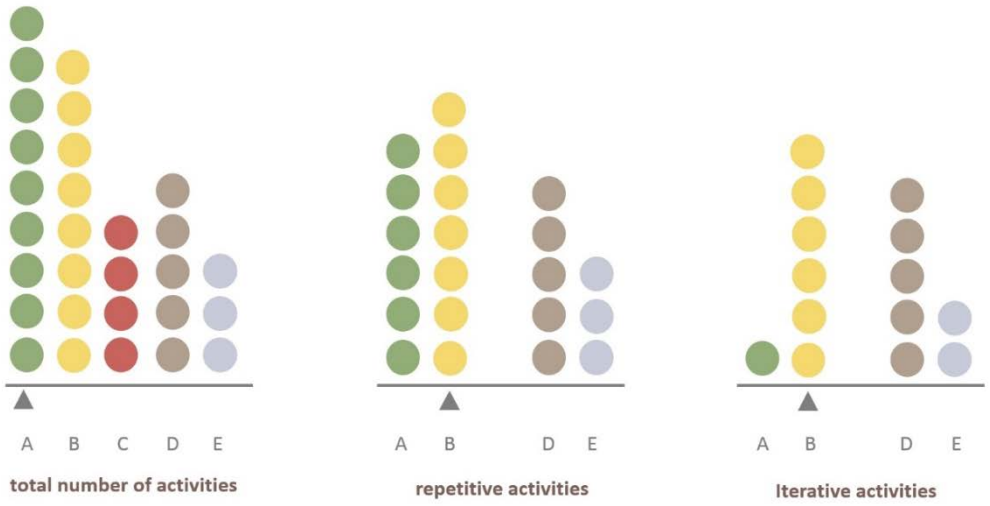


Fig. 8 Recurrence patterns inside the ordinal time visualisation of the production process of 3D virtual reconstructions of the old Town Hall of Krakow.
1 – no recurrences, 2 – only one type of recurrence during the activity (iteration or repetition), 3 – two levels of recurrent behaviour (iteration and repetition), 4 – high recurring motif (three and more)

Fig. 9 Recurrence patterns during the production process of 3D virtual reconstructions of the old Town Hall of Krakow represented using histograms.



The three histograms show the total number of activities within the process and the number of repetitive and iterative activities. Each histogram's bar corresponds to a given group of activities. A - data filtering and treatment, B- data analysis, C -data collection/acquisition, D - added value procedural activities, E – finalisation

[2] Predominant in the sense of the number of activities.

Concerning this case study, individual analysis does underline some interesting features, for instance, the strong contrast between the 3D modelling technical chain and previous sequences of activities. Activities belonging to the data analysis group, as well as added value procedural activities show a deep recurrent trend – they are often in the picture when iterative or repetitive sequences are present. Those related to the data filtering and treatment group are predominantly repetitive. Activities categorised into the data acquisition group do not follow this drift at all. The predominant group of activities in the process is the data filtering and treatment group (Fig. 9).
The visual analysis of an individual process helps us to understand some of its aspects (e.g., summarizes visually types of activities mobilised to produce an output). Additionally, comparing a process to other processes may reveal unthought-of tendencies and exceptions. It should be made clear that the Krakow Town Hall reconstruction case study was picked up to serve as a preliminary benchmark, a reference experiment against which further reconstructions processes may be compared. This is further commented on in the next section.

Comparative analysis

Visualisation can help characterizing a given element inside a collection, but it is often even more helpful when trying to derive knowledge from the collection as such, by cross-examining individuals inside the collection. D. Straker (2005) writes: "… *Comparative reasoning establishes the importance of something by comparing it against something else on the basis of an objective standard. The size of the gap between the things is used to infer some new assumption.* ...." In this manner, data visualization, considered as a cognitive activity (Spence 2001), can be of great help in spotting patterns inside collections.

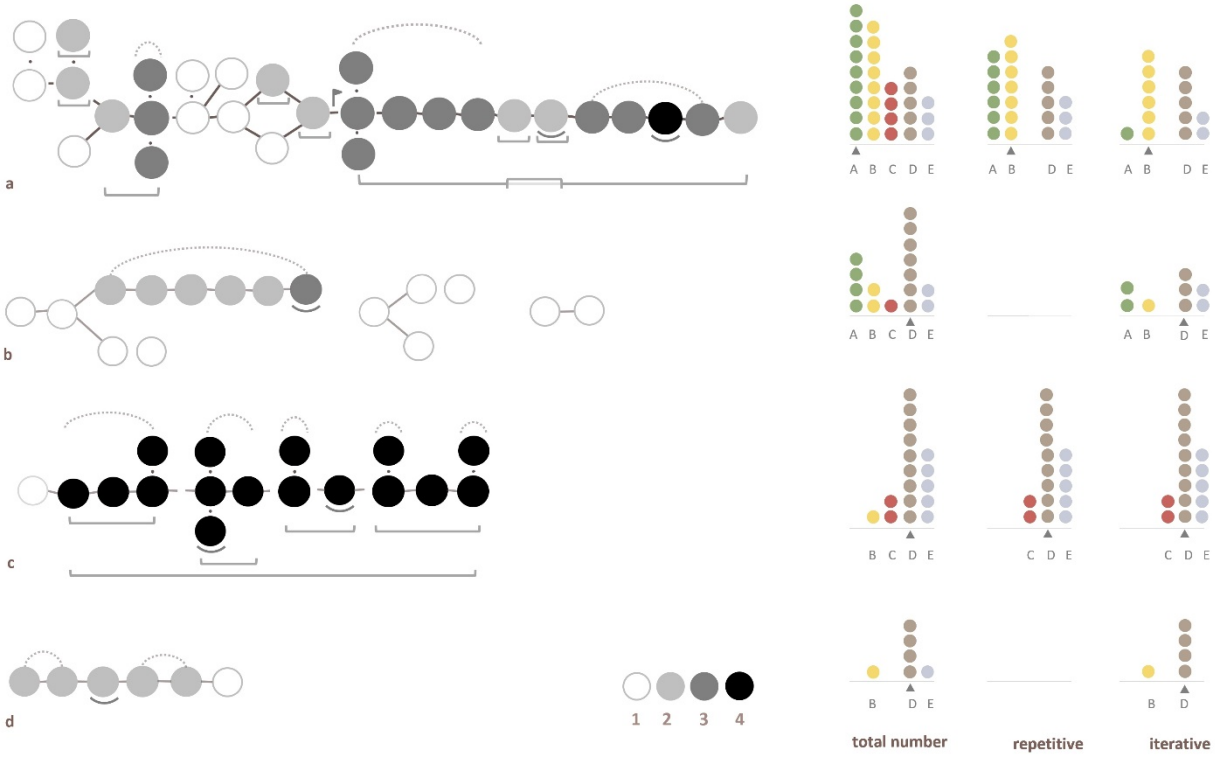


Fig. 10 A comparison of four processes using different visual formalisms and highlighting differences between these processes in terms of recurrence patterns.

Left – ordinal time visualisation; 1 – no recurrences, 2 – only one type of recurrence during the activity (iteration or repetition), 3 – two levels of recurrent behaviour (iteration and repetition), 4 – high recurring motif (three and more).
Right – histograms indicating the total number of activities within a process and number of repetitive and iterative activities. Each histogram's bar corresponds to a given group of activities.
A - data filtering and treatment, B- data analysis, C -data collection/acquisition, D - added value procedural activities, E – finalisation.

At this stage of the research, we have carried out a first experiment through which we try to weigh the potential of the approach in terms of comparative analysis. The process of production of hypothetical reconstructions of the Krakow Town Hall (a), is compared to three processes that led to three distinct outputs (Fig. 10): design and production of the *FlightSchedule profiler* proof-of-concept prototype (b) (Blaise & Dudek, 2016), production of PDF forms used topopulate the MEMORIA system in "off-line conditions" (c) and the design of one visual element of MEMORIA interface (d).
A comparison of those processes underlines strong differences in their structure, in the number of activities involved in each process, in the recurrence patterns. Concerning the latter feature, what is common for all of these processes is the predominance of iterative activities when observing those belonging to the 'added value procedural activities' group (D).
A common denominator for processes *b*, *c* and *d* is the predominance of the 'added value procedural activities' group of activities[3] (*e.g.*, programming, debugging, visual encoding, interaction design, testing). This may be a particularity of our team's know-how or way of doing, but to be true it is also an indication that we could be witnessing an epistemological shift. Indeed, it may be an indication that the impact on our everyday practices of *going digital* is far bigger than what we imagine.
However, that common denominator might stem also from the nature of the outputs. Does the characteristic of a process depend on its output, its authors, disciplines engaged or another factor? It is too early to formulate a plausible hypothesis. It is likely that what we will unveil are rather common traits delineating groups of processes, than general rules. We will need more data on more processes leading to more distinct outputs and preferably created in other research units, in a more interdisciplinary context. This is part of the future works planned, but at this stage what can be said is that the combination of a formalisation of processes and of visual solutions helping to analyse them already acts as food for thinking.

Implementation and future works

The project's initial development steps included a series of actions aimed at building a workable computer infrastructure composed of core elements such as an ontology of '*activities*' and their attributes, formalisms used to order and sequence '*activities*', selection of visual languages dedicated to the interfacing and candidate visualisations.
At the time of writing of this paper, the underlying data management system is operational, and the domain knowledge ontology is described. Yet the actual interfacing of the MEMORIA information system is still under development.
The development uses on one hand a relational database management system and on the other hand visual solutions based on JavaScript/SVG components embedded inside HTML pages. One of the particularities of our strategy in terms of implementation is the development of an

[3] A phase of research centred on the use of procedural knowledge, such as scientific procedures and technological protocols, and implicating the use of technical skills and abilities acquired and developed by training or practice.

off-line solution in parallel with classic online PHP-powered forms. The interface allows users to feed the information system using a set of predefined and downloadable PDF (Portable Document Format) forms that can be filled in at any time, with or without access to an Internet connection. Forms once they are filled in, can be sent to the server at any time for processing and storage. If the content uploaded into the system extends existing lists (typically, lists of creators, authors, institutions, etc.) new values are added into FDF (Acrobat Forms Data Format) files used to update lists inside the PDF forms.
There are two motivations behind this implementation choice:

- avoid putting researchers in a position where they *have to* be online,
- ensure that they keep trace of the information sent to a "remote" IS (PDF forms users fill in and upload to the MEMORIA system are stored on their own computer).

We have started to populate the system with real cases (some erstwhile outputs already treated, and over a dozen processes related to the MEMORIA system conception and construction). We are now in the process of describing on an everyday basis data acquisition processes conducted in the context of the Sesames ANR project[4]. The idea is to thoroughly investigate the approach's applicability and impact in terms of workload if used as a "day-to-day" registration tool.
Within the same formal framework (Sesames project) we intend to launch an experimentation on the case of alternative restitutions of the Abbey of Marmoutier (Tours, France), a case study that will be compared to this presented in this paper.

Final remarks

The MEMORIA project searches to comply with a logic of scientific integrity and good practices by experimenting practical solutions for the formalisation and description of research workflows. The initiative bases on the idea that beyond metadata describing outputs themselves, the scientific community concerned is awaiting means to ensure their verifiability, reproducibility and comparability.
Pursuing this long-term goal we develop an experimental information system aimed at empowering scientists in the field of heritage sciences with practical means to carry out the description, structuring, and analysis of their research processes. A set of activities leading to the production of an output is formalised as a '*process*', a concept whose role is to converge pieces of knowledge (*i.e.* data, information and relation between them) about research workflows.
One single process can be assessed, structured and described in several different manners. The choice of granularity of analysis to prevent information overload is one of the foremost issues. The final result will greatly depend on the capacities of scientists to model and represent their own method of work (*e.g.*, the consciousness of their methodological approach), on the importance they attach to intersubjectivity of the results they produce, on their will to be honest, etc.
In other words, a number of human and contextual factors will condition further steps, and more generally the applicability and added-value of the approach. Will they be distorting or contributing elements? Future works will shed a light on this 'shady valley'. With new opportunities often come new issues, new difficulties, and new hazards.

[4] UMR 3495 CNRS/MC MAP, 2019 ANR SESAMES project (Sémantisation Et Spatialisation d'Artefacts patrimoniaux Multi-Échelles : annotation 3D, Sonification et formalisation du raisonnement), Mai 2019. Available at http://anr-sesames.map.cnrs.fr/ [11 October 2019].

References

Aigner, W Miksh,S Schumann, H & Tominski C 2011 Visualisation of time-oriented data. London: Springer-Verlag.

Atkinson, M et all. 2017 Scientific Workflows: Past, Present and Future. Future Generation Computer Systems, 22 June 2017, [https://hal.archives-ouvertes.fr/hal-01544818/document l15 October 2019].

J.Y. Blaise, I. Dudek, 2016 The FlightSchedule Profiler: An Attempt to Synthetise Visually an Airport's Flight Offer in Time and Space. In: Fred, A et al. (ed.). Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR, Porto - Portugal, November 9 - 11, 2016. Porto: SciTePress, pp. 407-412.

I.Dudek, J.Y Blaise, 2017 What comes before a digital output? Eliciting and documenting Cultural Heritage research processes. International Journal of Culture and History, 3(1): 86-97. DOI: 10.18178/ijch.2017.3.1.083

Bocheński, J M 1968 The methods of contemporary thought. New York: Harper & Row.

Brewstera, C & O'Harab, K 2007 Knowledge representation with ontologies: Present challenges - Future possibilities. International Journal of Human-Computer Studies, 65(7): 563-568. DOI: 10.1016/j.ijhcs.2007.04.003

Burge, J E 1998 Knowledge Elicitation Tool Classification, 23 April 1998. Available at https://web.cs.wpi.edu/~jburge/thesis/kematrix.html#_Toc417957386 [10 October 2019].

CIDOC, 2017 CIDOC CRM. Available at http://www.cidoc-crm.org/ [Last accessed 10 October 2019].

CNRS, 2016, Integrity And Responsibility In Research Practices A Guide, CNRS. Available at http://www.cnrs.fr/comets/IMG/pdf/integrity_and_responsibility_in_research_practices_a_guide_05.12.16-2.pdf [Last accessed 01 October 2019].

Doerr, M 2003 The CIDOC CRM - an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 24(3):75-92. DOI: 10.1609/aimag.v24i3.1720

Doerr, M & LeBoeuf, P 2007 Modelling Intellectual Processes: The FRBR - CRM Harmonization. In: Thanos, C Borri, F and Candela, L (Eds.) Digital Libraries: Research and Development, First International DELOS Conference, Pisa, Italy, February 13-14, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer, pp. 10-14.

I. Dudek, J.Y. Blaise, L. De Luca, L. Bergerot, N. Renaudin, 2015 How Was This Done? An Attempt at Formalising and Memorising a Digital Asset's Making-of. In: Remondino, F et al. (Eds.) Proceedings of the 2015 Digital Heritage International Congress, Vol 2, Analysis & Interpretation Theory, Preservation & Standards Digital Heritage Projects & Applications. Granada. Spain: IEEE Computer Society, pp. 343-346.

Few, S 2014 Data Visualization for Human Perception. In: Lowgren, J et al. The Encyclopedia of Human-Computer Interaction, 2nd Ed. The Interaction Design Foundation. [online access at

https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception last accessed 10 October 2019].

Francis, J 2019 Tips for Creating an Effective Workflow Model, 10 January 2019. Available at https://kissflow.com/workflow/workflow-model-tips-for-creating-effective-workflow/ [Last accessed 15 October 2019].

Guercio, M & Carloni, C 2015 The research archives in the digital environment: the Sapienza Digital Library project. Italian Journal of Library, Archives, and Information Science, 6(1): 1-19. DOI: 10.4403/jlis.it-10989

Joliveau, T 2007 Echelle 1:1 et représentation grandeur nature, 30 April 2007. Available at https://mondegeonumerique.wordpress.com/2007/04/30/echelle-11-et-representation-grandeur-nature/ [Last accessed 09 September 2019].

Kotler, P et all. 2006 Marketing. Frenchs Forest, N.S.W: Pearson/Prentice Hall.

O'Hagan, A 2019 Expert Knowledge Elicitation: Subjective but Scientific. The American Statistician, 20 March 2019, [online access at https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1518265 last accessed 10 October 2019]. DOI: 10.1080/00031305.2018.1518265

Peltoniemi, P 2008 Is it possible to study scientific concepts, In: Madsen, BN & Thomsen HE (Eds.) Proceedings 8th International Conference on Terminology and Knowledge Engineering, Copenhagen: ISV, pp. 123-136.

Rosnay, MMD & Musiani, F 2012 The preservation of digital heritage: epistemological and legal reflections. Journal for Communication Studies, 5(2): 81-94.

Shadbolt, N R & Smart, P R 2015 Knowledge Elicitation: Methods, Tools and Techniques. In: Wilson, J R and Sharples, S (EDs.), Evaluation of Human Work, Fourth Edition, Boca Raton, Florid: CRC Press, pp. 163-200.

Spence, R 2001 InformationVisualisation. Essex, England: Pearson Education Limited & ACM Press.

Straker, D 2001 Comparative Reasoning, 30 April 2005. Available at http://changingminds.org/disciplines/argument/types_reasoning/comparison.htm [Last accessed 11 October 2019].

UMR 3495 CNRS/MC MAP, 2019 ANR Sesames project (Sémantisation Et Spatialisation d'Artefacts patrimoniaux Multi-Échelles : annotation 3D, Sonification et formalisation du raisonnement), Mai 2019. Available at http://anr-sesames.map.cnrs.fr/ [Last accessed 11 October 2019].