What Comes before a Digital Output? Eliciting and Documenting Cultural Heritage Research Processes

Iwona Dudek, Jean-Yves Blaise

What Comes *before* a Digital Output? Eliciting and Documenting Cultural Heritage Research Processes

I. Dudek and J. Y. Blaise

Abstract— **Knowledge-based systems, are today part of many research protocols where they act as powerful means to model, implement and cross-examine the workflows that lead from a set of inputs to a set of outputs.**

They remain however tricky to apply in the specific context of heritage science where workflows include a long tail of subjective human decisions, of non-explicit research protocols, of poorly formalised pieces of knowledge, of highly individual skills, of undocumented, non-reproducible, intuitive interpretations, when not simply of *licentia artistica*. Yet the heritage science community has witnessed over the past decades the emergence of huge quantities of digital outputs, either following massive digitization efforts, or as a result of the growing capacity of actors to produce digital-born material. How can this move be supported in terms of reproducibility, reusability and cross-examination of results if research protocols remain non-formalised one-shot efforts?

The research presented in this paper bases on the idea that what should be formalised and shared with future generations are not end results alone (outputs) but the methods and processes that lead the making of the output (human skills, tools, technological procedures, cognitive processes, scientific protocols, *etc.*).

Our contribution addresses a pending issue: how can we today complement traditional approaches to heritage assets documentation with means to describe and record research processes and workflows? The infrastructure we propose raises knowledge representation, visualisation, and information management issues. It applies primarily to a range of specific cultural heritage related artefacts, but is expected to be fairly generic in terms of methodology. In this paper we describe the methods employed in order to elicit underlying activities, support team elicitation through ad-hoc visualisations, promote a consistent visual interfacing of the underlying Information System.

Index Terms—**Information systems applications, knowledge extraction, elicitation and representation, visual reasoning, scientific protocols preservation.**

I. Introduction

In the past decades computer technologies and methods have little by little paved their way in the everyday practices of people engaged in the analysis and preservation of the Cultural Heritage (CH), becoming today simply central. This move came along with the production of huge quantities of digital outputs, either through massive digitization efforts, or as a result of the growing capacity of actors to produce digital-born material. The impact of computer technologies is not only visible across CH-related scientific disciplines, from for instance literary analysis to archaeological research, but also in very different application contexts such as communication, preservation, or data management. It is perceptible all along investigation processes, from for instance 3D surveying of heritage sites to the publishing of online repositories.

More data available, and more tools offered to analysts, can naturally be considered as creating new research opportunities, and in that sense contributes to the more general open science paradigm.

But the salience of computer-assisted investigation processes and workflows does also raise a number of methodological issues, among which these from which our research births:

1) Understanding and assessing the actual impact of the computer instrumentation on the research process and its outcomes, and notably differentiating clearly conscious choices made by the researcher from those, often unsaid, inducted by the technology at hand. Said briefly, *who drives the investigation - the analyst, or his computer-based "crutches"*? [1]
2) Communicating the whole process that lead to an output, to a conclusion, in other words letting others know how and why a certain result was achieved. As far as science is concerned, the reproducibility of an experiment or reasoning (i.e. being able to precisely describe the workflow that lead to a factual observation or to a general conclusion) is simply a cornerstone of any robust research process. Analysts must be given a chance to clarify and to communicate the choices they made all along the investigation – both in terms of technology and in terms of cognitive processes. [2]
3) Recording the way research processes are conducted, in order to spot chains of activities, better understand their respective weight, impact and frequency. Learning from

Manuscript received December 5, 2016; revised March 1, 2017. This work was supported in part by MCC, the French Ministry for Culture and Communication.

The authors are with the CNRS (National Body for Scientific Research) in the UMR CNRS/MCC 3495 MAP unit, Marseille, France (e-mail: iwona.dudek@ map.cnrs.fr, jean-yves.blaise@ map.cnrs.fr).

[1] For example, an archival text may mention that a building – today destroyed - was "two-storey high". The 3D modelling software an analyst will use in order to propose a virtual reconstruction of the building will not allow him to just say "two-storey": he will need to specify precise numerical dimensions. This situation somehow forces the introduction of a 'new' – and highly questionable – information into the research process. Such everyday life "small" impacts of technologies once added one to the other, may strongly impact the significance and interpretation of research results.

[2] A simple example of this situation can be found in the so-called "realistic" rendering of 3D scenes. Applying on the 3D objects textures that originate from a standard library of textures may be totally misleading for future analysts, for instance the size of bricks is likely in some geographical areas to be used as a marker of time – a "Baroque brick" differs from a "Romanesque brick". Applying on a 3D object the "wrong" brick, taken from a predefined library of textures, may over time drive future observers of the resulting 3D scene to erroneous conclusions.

doi: 10.18178/ijch.2017.3.1.083

experience means capitalising on what was done in order to refine what will be done. This can only occur if analysts are given a chance to register not only the achievement of an investigation process, but also the path leading to it.

The MEMORIA project is an attempt at complementing traditional approaches to heritage assets documentation with means to describe and record the research processes and workflows, including the choices made all along the process, and from which results stem. The expected outcome is an information infrastructure fostering (Figure 1):

1) *scientific reproducibility and cross-examination by associating an output to its creation process,*
2) *analysis of how research practices of individuals and organisations change over time.*

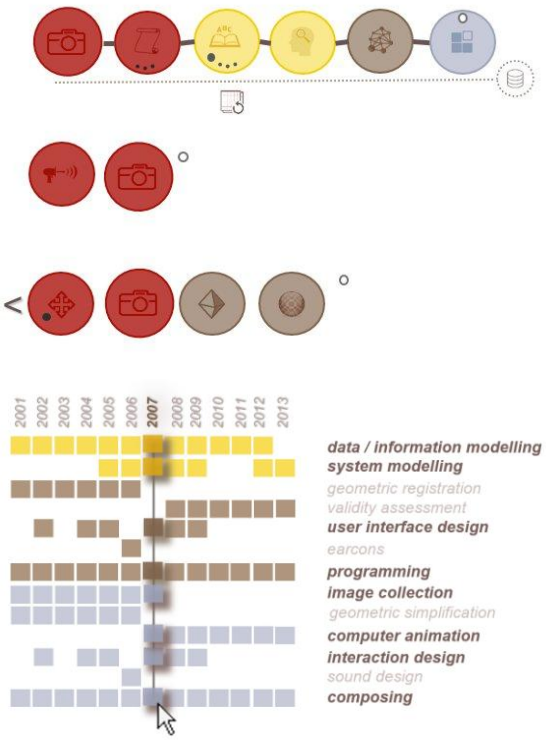

Fig. 1. Eliciting, visualising and recording workflows: top, processes seen as series of activities; bottom, evolution over time of research activities within a team.

In this paper we describe the methods employed in order to: (a) elicit the actual activities of scholars, (b) support team elicitation through ad-hoc visualisations, (c) promote a consistent visual interfacing of the underlying Information System (IS) that builds on universal visual metaphors.

In section 2 we position our contribution with regards to a general context – documentation of heritage assets – and with regards to the related challenges of knowledge and workflow elicitation.

Section 3 further discusses the research's objectives and sums up the concepts and methodological choices the project builds on (activities *vs.* process, inputs *vs.* outputs). Besides, the section contains definitions of the terms and notions used in the project.

Section 4 focuses on the identification and classification of activities, *i.e.* on the collaborative knowledge elicitation process. Section 5 discusses how that teamwork can be supported by knowledge visualisation solutions. Section 6 presents early "end-user" developments and introduces some key features of the forthcoming IS' visual interface.

It has to be said clearly that at this stage the approach targets a selected subset of CH investigation processes. The activities, workflows, processes that have been taken into account were limited to an area of concern that can be described as "*documenting historic architecture at various scales*" (ranging from furniture or statuary to whole sites or cities) through "all hints available" (ranging from textual inventories to dense matching surveys).

The necessity to enhance the "*readability and intelligibility of the scientific data and archives*" has been assessed by many, in particular in the digital Libraries community: Guercio and Carloni [1] for instance sum up some key risks for the persistency of digital archives and data. But efforts made to this day tend to focus on how to *describe the final result* (use of standards, addition of contextual information, specialized common vocabularies, interoperable schemas, *etc.*). The following statement can be found in Dulong de Rosnay end Musiani's epistemological and legal reflections [2] : "*[…] focus has often been put in recent years on legal and technical constraints concerning "the final output" rather on eliciting previous steps".*

Our contribution is precisely an attempt at overcoming this current pitfall, and what is more in the specific context of CH-related archives where the analysis process (pieces of knowledge and research protocols) tends to be quite poorly formalised. It introduces three main steps forward:

1) a step by step strategy for knowledge elicitation in strongly heterogeneous CH-related activities, and feedback from a real-life experiment,
2) a knowledge structure that has been built over time as a result of the above process, and that can at least act as food for thinking for researchers and collection holders,
3) a series of knowledge visualisation solutions that have proven efficient in supporting workgroup discussions, and that are fairly generic.

II. Research Context and Related Works

For some time now – one or two decades in fact – collection holders and other actors within the CH field have made very substantial efforts in digitising heritage assets. This effort is exemplified by various initiatives ranging from massive 3D scanning of artefacts to text digitization, or registration of old maps in geoportals. This move meets with two complementary needs: a need to preserve the heritage asset itself (may it be early cave paintings or a parchment), and a need for making it available in the context of the open science paradigm. Hence a number of online "libraries" (databases) are today made available for scientists as well as for the wide public, such as *Gallica* for literary sources, *Merimee* and *Palissy* for heritage artefacts (edifices and furniture), or *Europeana* at the European level.

Then came the time when born-digital material, started piling up in research labs. Often produced in the context of short-term communication actions, such outputs, although

they definitely can have a scientific value, tend to be considered as short-term consumer products – used once, then forgotten. [3]

Let's admit it: tomorrow's heritage assets will include what we produce today, born-digital material. Hence a number of initiatives at national or international level that have emerged in order to promote the preservation of digital assets (among which a notable example is naturally CIDOC, dedicated to the documentation of museum collections).

Shortly said, two issues have been addressed:

(1) Enabling semantic interoperability, *i.e.* developing common, robust and sharable description frameworks. (e.g. Dublin Core Metadata Element Set, CIDOC Conceptual Reference Model, AIM@SHAPE European initiative).

(2) Enabling technological interoperability and sharing end-user online tools. (e.g. COLLADA interchange file format, or through initiatives such as Culture 3D Cloud or 3DHOP [3]).

At the end of the day scholars in the CH field do have appropriate solutions today to capitalise, record and share research outputs. But our objective is to go one step beyond, or one step aside: *capitalise not only an output, but also a modus operandi*. We believe this step is vital, in the open science context, if we are to preserve digital-born heritage assets. Describing, preserving, sharing the "*how-to*" associated to a digital asset is a matter of science: not doing it would result in outputs piling up in research labs, with the research labs looking more and more like *Cabinets of Curiosities*. The issue was pointed out by M. Doerr and P. LeBoeuf [4] in a contribution where they introduce a realistic, explicit model of the intellectual creation process. This is basically what our research is about – only with a binding level of granularity in the description of activities.

In the context of the MEMORIA project what we wish to record corresponds to on the one hand the inputs to a study, on the other hand cognitive processes (e.g. inferences and choices made by an analyst, general knowledge applied) and processes mobilising procedural knowledge (e.g. individual skills, know-how, ways of doing). These processes are however, in real life, far from being homogeneous and equally well comprehended. We in fact face an application field in which formal procedures (may it be technical – a photogrammetric acquisition, or analytical – development of a specialised ontology) coexist with workflows relying on an individual's subjective choices (may they be technical - 3D modelling of complex shapes, or analytical – rating of credibility of historical evidence). As a consequence uncovering and describing these processes requires a thorny knowledge elicitation phase.

A. The Knowledge Modelling Issue

This research is about trying to characterize and record what people do in the course of a CH study. This includes a variety of distinct actions - complex ones, as well as trivial ones, long-lasting undertakings and short-lived efforts.

These activities share one thing though: they require human skills, human knowledge, and human decisions. Identifying and structuring them is often referred to in the scientific literature as *knowledge elicitation*: a process through which human activities are differentiated, named, described, and ultimately organised or modelled (often in the form of ontologies).

Bitter-Rijpkema, Martens and Jochems [5] propose a rather restrictive application of knowledge elicitation: … *how to support eliciting and sharing available but not yet articulated knowledge residing in the minds of individual team member.* … In this paper we tend to privilege a more general definition, this of Shadbolt and Smart [6] who relate knowledge elicitation to a statement of need: … *an attempt to elicit the knowledge of a domain expert* … But they further detail their view in a footnote: … *although early conceptualizations of knowledge elicitation cast the process as one of extracting or mining knowledge from the heads of experts, more recent conceptualizations view the process as a modelling exercise.* … This is precisely what is at stake in the MEMORIA initiative. Shadbolt and Smart [6] subdivide knowledge elicitation techniques into two groups: *natural* and *contrived* methods. The former correspond to informal situations such as interviews or the observation of actual problem solving, the latter to situation in which an expert undertakes a contrived task such as concept sorting.

Knowledge elicitation is a recurrent research topic in the context of managerial practices, as illustrated for instance by McInerney [7] or Yip and Lee [8], a context where the issue is, as worded by Gavrilova and Andreeva [9] … *to ensure that organizations extract as much value as possible from their knowledge*. … But what exactly would that *knowledge* be in our application context?

The challenging point in the MEMORIA project is that activities we wish to model and ultimately to record intermingle or encompass declarative knowledge and procedural knowledge. Whyte [10] provides the following definitions: declarative knowledge ... *involves knowing THAT something is the case* … (*e.g.* "I know that this semi-circular stone-built element is a round arch") whereas procedural knowledge … *involves knowing HOW to do something*… *(e.g.* "I know how to undertake a photogrammetric survey"). The difference is far from being anecdotal: CH-related studies most often imply the use of both these reasoning modalities, and whereas the former appears as falling into the scope of top-down knowledge representation approaches – ontology making typically – the latter is by essence harder to circumscribe and formalise. In section 4 we present a hybrid (natural AND contrived) approach to knowledge elicitation that we developed in order to cope with the strong heterogeneity of CH-related activities.

Another challenging aspect of the MEMORIA project is that in some cases activities will be carried out in a systematic, repetitive and well–defined way, but they can also leave quite a lot of room for on-the-spot improvisations – typically there are many different ways to produce a geometric shape such as a round arch in most 3D modelling software. Furthermore, an activity that appears as *a priori* well defined, can turn out to require stopgap improvisations – a car being parked alongside the object under scrutiny during a laser scanning campaign for instance. Recording each and every move, each and every action undertaken by an actor in the course of an activity may

[3] Prominent examples are 3D models, in particular virtual reconstructions: how many of these outputs, produced let's say between year 2000 and 2005, can still today be reused (or even just opened in a 3Dmodelling software)?

soon turn out to be a nightmare – a pointless ambition. In other words the knowledge modelling effort we report includes two sub-challenges:

(1) Building a knowledge model that pulls together well-formalised knowledge and specific, if not individual, know-how, skills, and ways of doing.

(2) Building a knowledge model with a granularity level in the definition of activities that avoids the cumbersome (and not necessarily significant) task of tracking and recording each and every action within an activity, and focuses on segregating significant, consistent steps in a CH study. As will be shown, differentiating the former from the latter is far from being straightforward – hence an in-depth collaborative knowledge elicitation phase presented in Section IV.

B. Workflow Management Issue

The production of research outputs most often results from a series of activities, *i.e.* from a workflow or process which is … *normally composed of tasks which are partially ordered*… [11]. Modelling and visualising this workflow can naturally be useful for error tracking, in quality assessment (precision of quantitative data for instance) or for communication purposes. But it can also be key in terms of productivity, of time-gaining. Understanding and formalising the chain of activities that lead to a given result allows to take up the study at any step all along the process and investigate alternative research paths – whoever may have conducted the initial study (*e.g. What would be the consequence of selecting alternative historical evidence? What opportunities of reuse of my colleague's 3D model as 3D prints?*).

In addition, on the long run, giving institutions means to record and compare workflows as they change over time can be greatly beneficial – typically helping to spot the most time consuming or recurrent activities, and thereby facilitating team management tasks.

There is vast literature, both in the Knowledge Management and in the Visual Analytics (VA) communities, on workflow modelling and visualisation. The range of application fields (including ad-hoc languages or tools) shows the issue is clearly frontier-less (see for instance [12] and [13]).

In the context of CH-related activities, there are many situations where a rigid calendar-based report of activities in linear time simply makes no sense. Some activities are inherently feedback-loop like (*e.g.* design and evaluation of graphic interfaces). Others – gathering of historical evidence for instance – may take months but with only a small effective time spent on the task since it is conducted as a sort-of episodic background task. In addition, in many cases, an actual ordering of activities (activity ***a*** before activity ***b***) would be misleading – sometimes several activities are conducted in parallel, and impact one another (3D modelling, analysis of historical evidence, experts giving their views, *etc*.). For all these reasons we consider that a realistic workflow management approach in our application field requires a high flexibility.

This challenge of facing "*imperfect knowledge*", as worded by Gershon [14], has been pickup up for instance by Aigner et al. [15] who introduce a solution called *PlanningLines* where time slots corresponding to a series of ordered activities are allowed to "slide" in time.

MEMORIA's approach to workflow management builds on complementary models of time as defined by Aigner et al [16] - *ordinal time* (***a*** occurs before ***b***, but neither ***a*** nor ***b*** need to be precisely known), *unanchored time* (an activity cannot be dated, or is recurrent), *branching time* (a process develops into two independent sub-processes), *etc*. The processes we need to formalise and record are NOT necessarily clearly anchored in time, nor are they always clearly ordered. As will be shown in Sections IV and V, the infrastructure we are building allows for the creation of processes (chains of activities) in which several alternative solutions are offered to date and put activities in relation with one another.

III. Background and Objectives

One thing is to come out with new, often born-digital, research results, but another thing is to make these results available and intersubjective in a way that allows for cross-examination, validation and reuse. In this section we first comment on the statement of need our research stems from. We then propose definitions for the main terms and notions that are at the heart of the approach in order to disambiguate the reading of the subsequent sections.

A. General Objectives — Statement of Need

The MEMORIA initiative aims at allowing the documentation of research processes and workflows that lead to the production of an output. We stated above in what this is a challenging issue in terms of knowledge elicitation. Let us exemplify what we consider should be seen as an emergent statement of need in the context of CH documentation practices. Fig. 2 shows a virtual (and hypothetical) reconstruction of the old Town Hall in Krakow. In a classic approach to CH asset documentation, this outcome would be associated with descriptors such as *author*, *creation date*, *file format*, *size*, *copyright data*, *etc*.

Fig. 2. A virtual reconstruction of the old Town Hall in Krakow (left - period 1686-1700, right - period 1454-1499). Note, for instance that a specific graphical encoding is used here to differentiate remaining parts (the tower, translucent dark-greyish colour) from the hypothetical parts (coloured or textured with more realistic way, opaque).

Are these descriptors enough for someone studying this same edifice today, fifteen years after the model's creation, to understand to which extent it is a trustworthy and reusable

hint? Do these descriptors tell us anything really significant about choices the authors made in terms of technological process (e.g. underlying survey techniques, structuring of the geometric model, texturing, …) or in terms of cognitive process (selection of historical evidence, justification of hypotheses, graphical encoding, …)?

And if these descriptors do not help in stating how authors came to produce such an outcome, in what is it, in terms of scientific significance, useful at all for future research efforts?

And so the problem is simple: *what would be needed if this outcome was to serve as a robust, reusable input in subsequent research processes*? In this specific example the answer, intermingling technological and cognitive processes, would be for instance making sure that the following information set is recorded:

- What kind of geometric data is used? Are dimension based on a survey process, on someone else's hypothesis, or on a historical inventory?
- Is there a specific graphical encoding? If so, to what semantics do differences in the appearance of objects correspond?
- Basing on what historical evidence is the reconstruction proposed?
- Was the analyst faced with contradictory evidence? If so, what choices did he make, and why?
- When facing information lacks, to which extent did the author introduce analogies?
- If analogies were used, on what criteria do they base? (stylistic affiliation, regional specificities, time slot, *etc*.)

Only given these pieces of information will the above research outcome serve future knowledge making efforts – otherwise we will just look at it like Vivant Denon looked at hieroglyphs: observing them with enjoyment, but without comprehension.

B. Terms and notions

Actors engaged in collaborative normalisation or documentation efforts know very well that building a terminology – meaning here the way people name things and notions – often goes along with potential misunderstandings (double meanings, polysemy …). The same can happen when reporting on a research. Accordingly we have chosen to list and define in this section a certain number of terms and notions that are extensively used in our research (see also Figure 3).

TABLE I: THE MEMORIA'S INFORMATION ARCHITECTURE MAIN NOTIONS

term	***definition***	***example***
activity	An activity identifies a series of actions undertaken in order to produce resources all along a project's workflow. When the resource is to be described and stored in the Memoria infrastructure *[this is an analyst's decision]*, this resource is called an *output*.	*e.g. imaging, phonological disambiguation, data conversion, graphical composing*
output	A resource, usually digital or digital-born, resulting from one or more activities stored in form of a process. An output can be a simple document (a screenshot, a 3D model, a video, *etc*.) or a set of documents (a collection of models, charts, *etc*.). Categories of outputs are differentiated through a set of classes defining *types* and *subtypes*. An output can be associated with one or more objects of study. Outputs reused in subsequent activities are named '*inputs'*.	*e.g. a screenshot, a 3D model, a video, a collection of models, a collections of charts*
process	The notion of process is used to represent the chain of activities mobilised to produce an *output*, a *publication* or a *composition*. A process may include only one activity. The order of execution of the various activities within a process can be specified or left unspecified. The same activity can contribute to several independent processes. Each process can be linked to the preceding process in order to establish consistent and reusable sequences of activities in the framework of 'process template'.	*e.g. successive steps within a survey campaign, that lead to the production of a 3D point cloud.*
input	An *activity* can be based on one or more sources (external resources), as well as it can be based on one or more *outputs* produced previously and described by a separate *process*. To avoid confusion the *outputs* used in an activity as initial root elements will be called *inputs*.	*e.g. a 3D reconstruction of a Greek temple produced previously, reused in a 3D printing process*
object of study	Any natural or manufactured object of heritage significance that has been studied and represented in an *output*. Each object of study is designated by one of eleven categories corresponding to the concept of scale (movable objects, earthworks, grottos and mines, architectural components, urban fabric, …). Objects of study are characterised by some basic taxonomical and geographical information, as well as temporal indicators, allowing to query the system on space+time search criteria.	*e.g. ensemble of edifices of the Main Market Square in Krakow, Pont Saint-Bénézet – Avignon, pieces of furniture of the Trianon, Versailles, etc.*

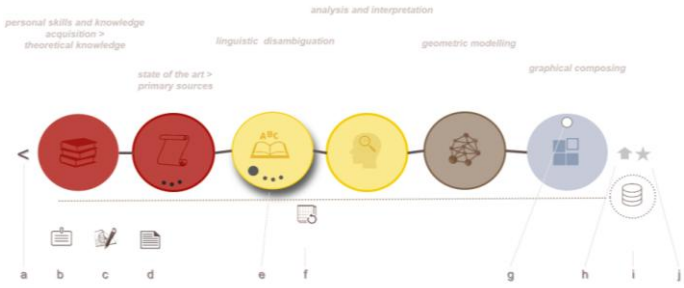

Fig. 3. Illustration of how the above notions get combined on the example of an ordered process based on the results of a preceding process and using an external infrastructure. Three types of results of this process were recorded: output (***g***), publication (***h***) and composition (***j***).
a –preceding process, ***b*** - process name, ***c*** - project in which the current process has been conducted, ***d*** – comments on the process, ***e*** - recurring activity, ***f*** - the process has led to a modification of the structure of the infrastructure that has been engaged in the process, ***g*** - link to the outputs resulting from the process, ***h*** - link to the publications resulting from the process, ***i*** - information on the infrastructure that has been engaged in the process, ***j*** - link to the compositions resulting from the process,

IV. Identifying and Structuring Activities

The MEMORIA initiative builds on the idea that *activities* and *processes* that lead to a given *outcome* can be identified (i.e. named, described through attributes when needed) and structured (i.e. organised with regards to an ontological approach when possible, combined to form chains when needed). Our approach can be described as standing in between two Knowledge Modelling (KM) practices: *knowledge elicitation* (spotting in existing workflows formal procedures and methods) and *reverse engineering* (digging out protocols and know-how). In the following sub-sections we first present the method used in order to identify, disambiguate and structure activities - *i.e.* the knowledge elicitation step. We then introduce and illustrate the resulting knowledge structure.

A. Identification and Disambiguation of Activities: A Collaborative Approach?

Shadbolt and Smart [6] comment on the knowledge elicitation process as follows: *…the idea is that the knowledge elicitor and domain expert work together in order to create a model of an expert's knowledge. …* We build on that common sense assertion, and carry out the task in a collaborative manner. Team members are pulled together in order to discuss the definitions and organisation of activities, with regards to their specific skills. But just pulling together people around a table and having them "say their word" can be quite time consuming, and relatively inefficient, if the discussion's scenario and ultimate goal are not made clear straight away.

Our approach is definitely a collaborative approach, but not only. It can be seen as loosely comparable to Peltoniemi's *concept disambiguation strategy* [17]. This author introduces a concept analysis method, originating from terminology work, and applied to the study of scientific concepts and methods. The statement of need he bases on is quite close to a central issue we are facing in our research when trying to identify and describe research workflows: *do specialists refer to the same concept when using one same term, or terms considered as close*?

Peltoniemi proposes a concept disambiguation method that is structured in successive steps:

4) concept/term extraction,
5) concept analysis,
6) representation of concept diagrams as an elicitation teamwork supporting tool,
7) definition of concepts,
8) evaluating and choosing adequate terms,
9) choosing equivalents.

These are broadly speaking the steps we went through in our research - only, in our case, the approach targets the analysis of activities and not of scientific concepts. In this section we present our implementation of these steps, except the representation aspect, discussed in Section V.

B. Extraction and Analysis of Activities

In the initial concept/term extraction step what needs to be done is on one hand identifying activities and on the other hand picking up terms and preliminary definitions from the existing scientific literature, as well as from the existing work practices in order to provide verbal representations of activities and research protocols (Figure 4). This implies selecting terms used to loosely describe an activity (*e.g.* "remote sensing") as well as terms that are related to a given result (*e.g.* a "3D cloud of point"). This was done at the very beginning of the project through workgroup meetings where actors were asked to list activities they had conducted or witnessed. At this stage, unlike in strict top-down concept extraction processes, both terms that do correspond to activities or processes and terms corresponding to a very operational level (for instance the name of a given, specific software - 3D model created with Blender) were extracted.

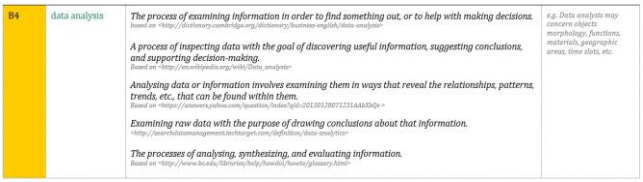

Fig. 4. An illustration of the concept extraction and analysis steps: spotting a notion/term and harvesting definitions.

In the second step, the concept analysis step, a significant effort had to be carried out in order to filter the list of terms and notions, and to interpret and assess relations between activities at the conceptual level (e.g. "remote sensing") and practical, down-to-earth activities (e.g. "I applied textures extracted from a set of photographs taken on the site to the 3D model in Blender"). Relations between activities were assessed, and activities belonging to a same "area of concern" were grouped and organised in several hierarchies (*e.g.* all data acquisition activities, ranging from on-site 3D surveying activities to the gathering of archival material). This step implied clarifying the *intension* of each *activity* - *i.e.* associating individual activities with diverse definitions in order to elucidate the internal content of a concept and assess its potential relations to subordinate / superordinate categories.

Each activity contains a certain numbers of descriptors that can be specific to the activity (classic refinement process), or not (Figure 5). This is because in the MEMORIA approach we sometimes faced situations when we needed to decide whether seemingly not-that-different activities were different or not (e.g. acquiring on-site photographs for documentary purposes or acquiring photographs in the context of a photogrammetric acquisition).

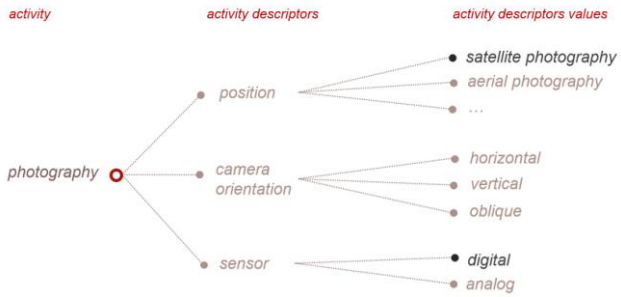

Fig. 5. An activity can be characterized by a number of descriptors that describe different aspects of this activity and make the definition narrower, or more specific. Several descriptors can be used to characterize one activity. Each descriptor may accept several values that may be defined as single choice or multiple choice field.

We combine an *intension-based* approach, with an approach *by list of properties*, to quote U. Eco [18] so as to clearly unveil situations where two similar activities containing the same attributes, need to be disambiguated.

At the end of the day the concept analysis step unfolds in three sub-steps:

1) giving verbal definitions of each activity,
2) organising relations between activities,
3) tracking errors by further investigating the content (list of attributes) of each activity.

The result of the extraction and analysis steps is a first, yet to be evaluated, ontological structure, built by combining a *top-down* approach (groups of activities) and a *bottom-up* approach (relating activities to practical skills, and technological environments). This structure is represented through graphics called *wheels of activities* (Fig. 6) that aim at clarifying the position of each activity in the hierarchy and at supporting the subsequent workgroup discussions (see also Section V).

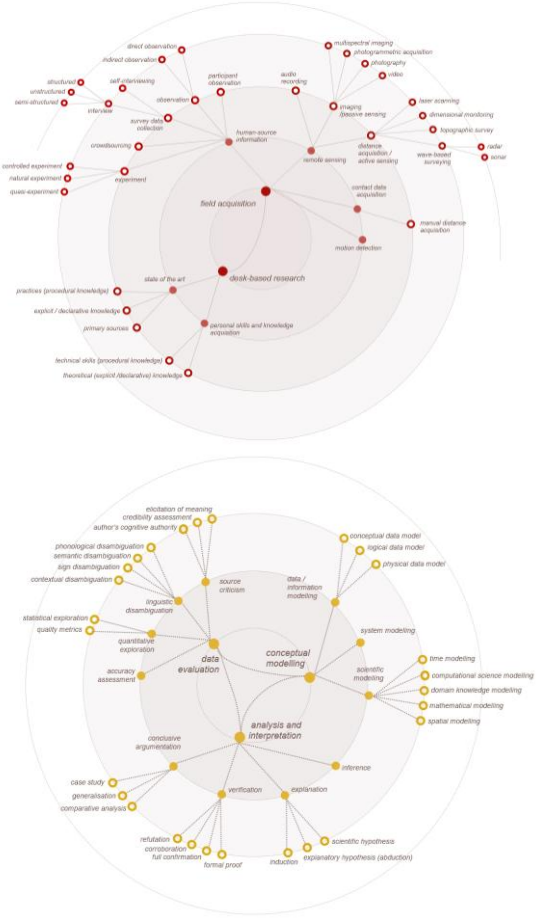

Fig. 6. The "wheel of activities": a visualisation of the hierarchical structures, applied here to two distinct groups of activities. Top - the data acquisition group, right - the data analysis group of activities. The most general notions are those closest to the centre of the wheel.

C. From Several to One: Definition, Terms, Equivalents and Examples

As already mentioned, at start several definitions are gathered from written or online literature for each activity. Once a preliminary ontological structure has been built and visualised, a series of steps are undertaken in order to further circumscribe activities by giving each of them a unique definition, a unique name, and by illustrating them by some concrete examples.

Knowledge elicitors and small groups of participants meet for half a day work sessions dedicated to groups of activities or skills. Together actors browse down a hierarchy of activities and discuss alternative definitions given to an activity, rephrase or combine definitions until a consensus is reached. Examples are commented as well, and in many cases the participants are asked to contribute by adding a *new*, or *better* example that would match the definition of the activity. This exemplification of an activity by participants, also called *ostensive definition* [19] is often absolutely crucial in the way they acquire a personal understanding of what the activity encompasses in terms of real-life workflows. When needed terms that are picked up as "equivalents" are quoted or disambiguated in the definitions.

As a result a final ontological structure is formed and visualised. It has to be said though that the finely ordered series of steps proposed by P. Peltoniemi acts more like a general framework than like rigid how-to-do instructions. In our "real-life experience" the process has turned out to be deeply iterative, with a number of feedback loops all along the building of the ontological structure through which the supposedly initial extraction step can be questioned and revisited at any time.

D. Using the Classification of Activities

The main result of the above steps is a classification of activities into five main groups organised hierarchically (altogether this represents 232 activities and 398 attributes so far, with three groups exhaustively formalised, and two partly). Table II lists the top categories of the resulting ontological structure.

TABLE II: TOP CATEGORIES

A	data collection	Gathering data which is to become the subject of further analysis, filtering and processing. *unfolds in: field acquisition, desk-based research*
B	data treatment	Transforming the raw data into a suitable form with regards to analysis, output production or finalisation needs, either when accessing the data for the first time, or in subsequent data steps. Editing, cleaning or modifying the raw data results in processed data. *unfolds in: image processing, geometric reconstruction, initial data manipulation, geometric re-processing*
C	data analysis	Methods of acquisition or gaining of scientific – theoretical, explicit – knowledge, as well as manners of its articulation and transmission in a formal language. *unfolds in: data evaluation, conceptual modelling, analysis and interpretation*
D	added-value procedural activities	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. *unfolds in: visual outputs production, user interface design, tangible outputs production, auditory outputs production, validity assessment, automation, image analysis, programming*
E	finalisation	Modification, adjustment, reformation, optimisation and adaptation of already produced outputs to particular needs *unfolds in: sound design, optimisation, composing, make-up, post-treatments, interaction design, animation, imaging*

One of the essential features of the system we build is to ensure its flexibility - minimize the amount of mandatory information, allow to alter the order in which information is input. For example:

1) The notion of *process* encompasses situations where the chaining of *activities* is clearly stated (ordered processes) and situations where activities cannot be ordered and are stored as a set of activities.
2) When a detailed description of each activity inside a production process is out of reach (typically when documenting outputs produced some time ago, or by some other team members), using superordinate activities allows users to depict the activity with broad strokes (e.g. mentioning only a remote sensing activity

when the origin of a 3D point cloud is not established).

3) Descriptors of activities allow for a fine-grain description of activities, but they most often are optional - users are not asked to make guesses.
4) Iterative processes or activities can be recorded as such (e.g. trial and error processes in the design of an IS interface for instance).

Altogether the idea is to allow for a level of flexibility in the recording and documentation of research processes and workflows, typically when needing to record processes carried out some time before its recording, by members of staff who moved. We consider it is vital if the approach is to be workable, to view the classification of activities as a tool more than as a normalisation framework.

V. Knowledge Visualisation Solutions

An important aspect of this research is the role given to *visualisation* seen as a mean to support the knowledge elicitation process. There is nothing particularly new though in considering this aspect as important.

In a collation of expert opinions [20] V. Sabol pinpoints the potential benefits of the knowledge visualization effort: … *Knowledge in visual form not only facilitates remembering and transfer, it also provides the fuel for reasoning processes where new knowledge is derived and created from previously acquired knowledge.* …

Supporting teamwork through visual means is in fact a recurrent research topic inside the knowledge visualization field, i.e. a field focusing on the ... *transfer of knowledge among persons* …, and working on … *smaller but highly organized sets of information* … [21].

Peltoniemi's concept disambiguation method [17] includes a step called '*representation of concept diagrams as an elicitation / teamwork supporting tool'* for which the author implements a so-called satellite system (a system where the most important concept is placed in a central node with others forming satellite nodes around it). The satellite system is used as a tool in the disambiguation process: visualising it helps giving an overview of the relations between concepts and supporting a progressive fine-tuning process. Although we introduce a different visual solution, we do base on this statement of need. In the context of the MEMORIA project we primarily focused on supporting the knowledge elicitation step, i.e. the building of an intersubjective and consensual ontological structure representing activities.

There is a vast literature on how to visualise hierarchical structures at large, ranging for instance from *hyperbolic browsers* [22] to *node-link diagrams* [23]. There are today a number of graph visualization software available on the net – the open source *Gephi* graph vizualisation platform [24] is probably the most widespread, on which we could potentially have based. But along with *a-priori* positive aspects of such "ready-to-use" solutions – efficient browsing and interactivity, comes what we have considered as major drawbacks: a steep learning curve on one hand, and a significant impact of the above mentioned browsing and interaction on the visualisation's cognitive load.

E.R Tufte [25] has worded a key principle of graphic design as follows: … *graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space*. … In line with this view we have developed a visual solution, built on design principles that foster a clear-cut, immediate, technology-free capturing of the underlying semantics. These principles can be summed up as follows:

1) a general visual organisation using a simple wheel-like concentric distribution of activities where hierarchical relations can be straightforwardly read, and alternative densities assessed visually at a glance,
2) a context view where all elements inside a given group of activities are presented, and focus views on sub-parts selected by the knowledge elicitor (Fig. 7, bottom),
3) a same visual organisation is used to represent "trees" of activities, and descriptors of activities (Fig. 8, bottom),
4) a very limited number of graphic entities, facilitating the decoding and memorisation (circles, lines),
5) nodes and groups of nodes can be repositioned or restructured with keeping a systematic overall organisation,
6) a colour coding of nodes (circles) is used to differentiate groups of activities (Fig. 6, Table II),
7) inside one group, during the elicitation process, a colour coding of texts (labels of nodes) and nodes is used to differentiate notions already discussed by the participants from those that remain either non-processed or non-consensual (Fig. 7),

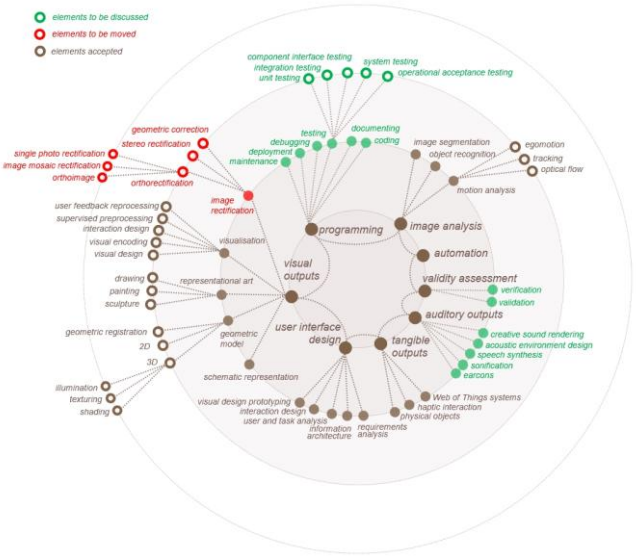

Fig. 7. The "wheel of activities" visualisation corresponding the added-value procedural activities. Colouring of the nodes and labels during the elicitation phase is used to distinguish elements depending on the level of completion of the disambiguation and definition tasks.

8) a reuse of the graphic's concentric organisation, developed in order to show "at concept level" how groups of activities unfold in subordinate activities and to assess relations between activities, in the representation "at instance level" of a specific process or the whole activities carried in a given project for example (Figure 9).

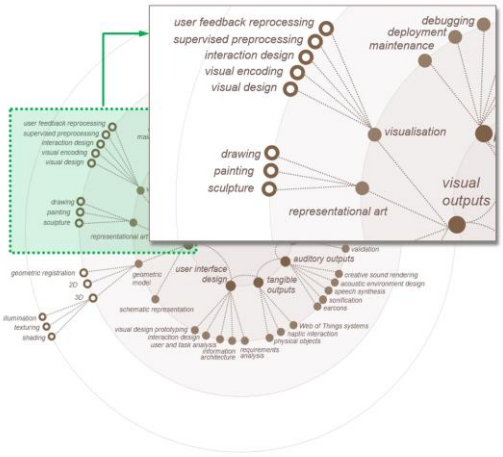

Fig. 8. The "wheel of activities" visualisation corresponding the added-value procedural activities. The visualisation allows a general context view on a group of activities, and user-chosen focus views on subordinate activities (schema).

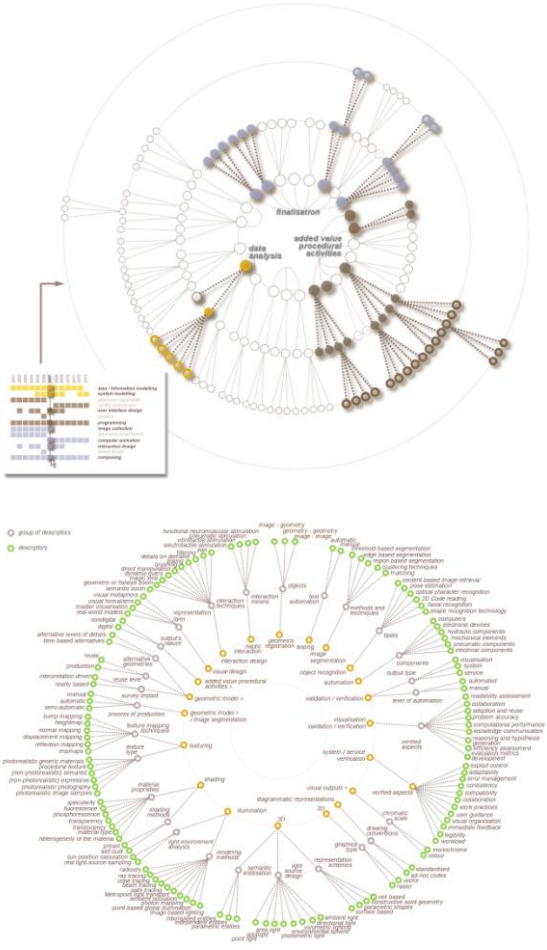

Fig. 9. The graphic's concentric organisation is reused in (top) instance-level representations of a specific process or of a team's yearly activity and in (bottom) the concept-level representation of relations between activities and their descriptors.

As illustrated above, our approach can be seen as building on the concept of simplicity in design and technology as defined by [28] and discussed in [27] or [28].

What can be said at the end of the day is that these visualisations greatly facilitated the workgroup discussions. Typically, when facing textual inputs (tables as shown in section 4.2), participants repeatedly turned to the visualisation in order to keep a clear understanding of the overall hierarchical structure and check whether notions were not duplicated or ill-positioned in the hierarchy. Shortly said, the visualisation helped reducing the level of verbal thinking required from participants in order to read the relations between concepts, remember names, identify missing concepts, *etc*.

VI. Current Developments

In this paper we have chosen to put a strong focus on the knowledge elicitation and visualisation aspect of the MEMORIA initiative, since we consider they are key building blocks in any knowledge modelling and documentation effort. But the project's ultimate goal is to implement a web-enabled IS, allowing actors in the CH field to memorise research processes and workflows through everyday interactions with the IS. The implementation being developed basically combines an RDBMS and web interfaces.

It has to be stated clearly at this stage that the IS development is still at an intermediate stage: the knowledge structure and its interfacing (Figure 10) have been tested so far only on parts of the overall ensemble of categories. We have therefore not reached a point where we could conduct an in-depth analysis of impact and usability. Yet there are key choices that have been made in terms of implementation, and that can give the reader a good idea of how the approach will be mapped in to a concrete, real life computer platform. In the following sub-section we briefly comment on two prominent design choices: scenarios of use and visual language.

A. The IS Usage Scenario

We have put quite a lot of emphasis on the notions of *activities* and *processes* throughout this contribution. But the scenario of use we privilege remains closer to classic documentation efforts in the sense that all starts by the recording of an *output*. The interface is designed to allow, at any time during the research workflow, an actor to insert an output and initiate a new recording process.

The recording process starts by the identification of an output considered as worth recording and storing. The output can be a single document (a 3D model for instance) or a set of documents (e.g. a whole web site). Subsequently several steps can be combined:

1) identification of activities that were conducted,
2) linking of activities to form ordered or non–ordered processes,
3) spotting of iterative chains of activities inside the process,
4) reference to an underlying infrastructure,
5) reference to a previous process,
6) listing of inputs and external sources needed as a prerequisite in carrying out the activities,
7) reference to external interventions (expertise),
8) description of the institutions, projects, actors, objects of study,
9) description of the publications and compositions in which the output is used or reused,
10) extension of the 24 controlled vocabulary lists when needed, *etc*.

Fig. 10. Screenshot of the IS interface in search mode.

There are naturally two major ways to "go back in time" in order to record all successive activities: one is to start by the very beginning (e.g. "*I started by converting data I got from my partner and then …*") or the other way round, in a sort-of reverse engineering approach (e.g. "*just before I produced this output I was working on the texturing, and before that …"*). The recording process is basically user-dependant: the above steps can be freely intermingled (in the limits of information dependences naturally). A personal "working zone" is attributed to each contributor in order to further facilitate a progressive and flexible insertion of data.

The usage scenario privileges an *a posteriori* recording of activities, once an output has been produced. Naturally another option would be to try and track each and every move "on the fly" - we consider this option would not only as over-invasive, but is simply irrelevant in a research context where workflows include highly individual decisions and know-how, and can include loose trial and errors protocols that go beyond the project's scope. The MEMORIA project is not about tracking everyday activities of a staff, it is about making sure future generations can reuse the scientific output we are today producing – and this makes a difference.

B. The Visual Language: Metaphors and Formalisms

We consider a crucial aspect in the MEMORIA initiative is developing a *visual interface* that provides on one hand access to the results of queries on all outputs (sorted by object of study, project, production process, *etc.*) and on the other hand that shows the evolution of methods, techniques and tools used over time, as well as of types of activities. What is meant in this section by *visual language* is the result of a specific effort that designers of IS can make in order to ensure consistency between on one hand the semantics behind the IS and on the other hand the modalities offered to users in terms of interaction.

Our first priority has been the design of a *visual metaphor*, building on the image of an analogue film that conveys a variety of information about the output: type sizing, authors (*etc*.) in a synthetic (and wordless) way. This visual metaphor is used to sum up and communicate visually the notions behind the approach and the state of completion of the documentation effort. The film metaphor marks the visual identity of the project, and is reused for most of the other components of the system - composition, object of study, project, *etc*. (Fig. 10, 11).

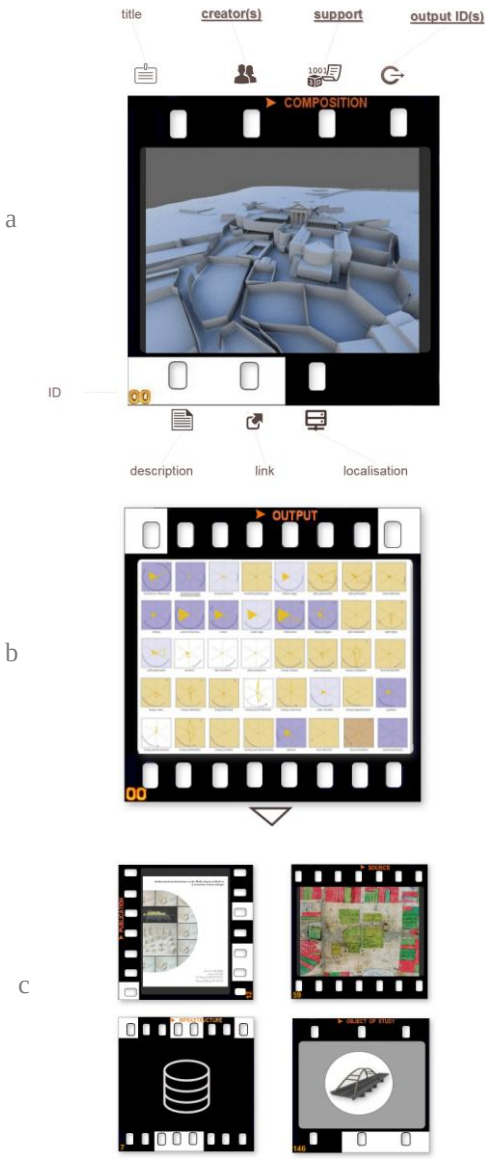

Fig. 11. Several uses of the "film" metaphor to sum up the completion of the documentation effort for (a) a *composition*, (b) an *output*, (c) other components: a *publication*, a *source*, an *infrastructure*, an *object of study*. Film perforations correspond to the attributes of a component – the "perf" appears on a white background until the attribute is filled in by the contributor. (a) An onmouseover event on each perf triggers the opening of a textual description for each attribute of a component, along with corresponding icons. (b) the' down arrow' is left unfilled, underlining the fact that the *output* has not yet been associated to a *process*.

But the film metaphor only conveys information about the level of completion reached in the description of outputs, projects, expertise, compositions, publications, objects of study, infrastructures, and sources - not about the activities as such.

Activities and *processes* are visualised through simple glyphs that reuse the colour coding mentioned in table II (corresponding to categories of activities). A set of symbols complete the graphic vocabulary associated to the representation of chains of activities so as to spot iterative chains, sub-categories, *etc*. (see Fig. 3). Additional visual formalisms are introduced in order to represent other pieces of information such as teams and individuals (Fig. 12).

Fig. 12. Visual formalisms for institutions and individuals – note the reusing of the colour coding used in order to distinguish categories of activities.

VII. Conclusion and Perspectives

The MEMORIA research aims at allowing CH scholars to document and record, beyond a research result, the process and workflow that lead to that result. In this contribution we have reported on the method implemented in order to carry out what we view as a crucial aspect of such an approach: *knowledge elicitation* and its *support by visual means*. Two significant limitations have to be mentioned in order to weigh the potential impact of the MEMORIA research at this stage:

1) A full-scale implementation of the approach, in terms of computer platform, has not yet been reached: gaining insights on how it can be deployed across a variety of use cases requires further investigation.
2) The approach is expected to be fairly generic – insofar as the knowledge elicitation phase in concerned – but up to now we have tested it only on a limited subset of activities.

Given this, we do believe that at a time when digitized or digital-born outputs are massively produced, the scientific community needs further exchanges of ideas and experiments in order to make of these massive data and information sets an opportunity rather than a burden, an open challenge rather than a hidden side-effect. In that sense, we hope our research can contribute to pinpointing some clear emerging research challenges:

1) better understanding the impact of the computer instrumentation on today's research processes,
2) fostering scientific reproducibility and cross-examination,
3) recording the way research processes are conducted in order to spot chains of activities, and better understand their respective weight, impact, and frequency,
4) e for the analysis of how research practices of individuals and organisations change over time in terms of methods of work, activities and competences, thematic mobility (objects of study, cooperation, expertise required), and so on.

This still ongoing research illustrates what we consider is one of the major challenges the heritage science community will have to face in the coming years: sensemaking in massive data sets.

References

[1] M. Guercio and C. Carloni "The research archives in the digital environment: the Sapienza Digital Library project," *JLIS.it* ,vol. 6, no. 1, pp. 1-19.

[2] M. D. D. Rosnay and F. Musiani, "The preservation of digital heritage: epistemological and legal reflections," *Journal for Communication Studies*, vol. 5, no. 2, 2012, pp. 81- 94.

[3] M. Potenziani, M. Callieri, M. Dellepiane, M. Corsini, F. Ponchio, and R. Scopigno, "3DHOP: 3D heritage online presenter," *Computer & Graphics*, vol. 52, pp. 129-141, November 2015.

[4] M. Doerr and P. LeBoeuf, "Modelling Intellectual Processes: The FRBR - CRM Harmonization," in *Proc. ICCOM-CIDOC Annual Meeting Conference*, 2006. pp. 10-14, 2016.

[5] M. Bitter-Rijpkema, R. Martens, and W. Jochems, "Supporting knowledge elicitation for learning in virtual teams," *Journal of Educational Technology and Society*, vol. 5, no. 2, pp. 113-118, January 2002.

[6] N. R. Shadbolt and P. R. Smart, "Knowledge elicitation," in *Evaluation of Human Work*, J. R. Wilson & S. Sharples, Eds., Boca Raton: CRC Press, 2015, ch.7, pp.163-200.

[7] C. McInerney, "Knowledge management and the dynamic nature of knowledge," *Journal of the American Society for Information Science & Technology*, vol. 53, pp. 1009–1018, 2002.

[8] J. Yip and R. Lee, "Knowledge elicitation practices for organizational development intervention," *Knowledge Management Research & Practice*, Palgrave Macmillan, January 2016.

[9] T. Gavrilova and T. Andreeva, "Knowledge elicitation techniques in a knowledge management context," *Journal of Knowledge Management*, vol. 16, no. 4, pp.523–537, 2012.

[10] S. Whyte. (April 2016). Declarative vs Procedural Knowledge. [Online]. Available: http://unt.unice.fr/uoh/learn_teach_FL/affiche_theorie.php?id_concept=90&lang=eng&id_theorie=1&id_categorie=3

[11] Y. Yang, W. Lai, J. Shen, X. Huang, J. Yan, and L. Setiawan, "Effective visualisation of workflow enactment," in *Proc. the 6th Asia-Pacific Web Conference, Advanced Web Technologies and Applications*, eds. J. Xu Yu, X. Lin, X. Lu and Y. Zhang, 2004, pp.794–803.

[12] R. Brown and H.Y. Paik. "Multi-faceted visualisation of worklists," in *Journal on Data Semantics XII*, vol. 5480 LNCS, ed. S. Spaccapietra, pp. 153 - 178, vol. 5480 LNCS, 2009.

[13] M. Radzikowska, S. Ruecker, G. Rockwell, S. Brown, and L. Frizzera. (April 2012). Workflows as Structured Surfaces. *Digital Humanities 2012.* Available: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/workflows-as-structured-surfaces.1.html

[14] N. Gershon, "Visualization of an imperfect world," *IEEE Computer Graphics and Applications*, vol. 18, no, 4, 1998, pp. 43–45.

[15] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl. "PlanningLines: novel glyphs for representing temporal uncertainties and their evaluation," in *Proc. the 9th International Conference on Information Visualisation (IV'05)*, 2005, pp. 457 – 463.

[16] W. Aigner, S. Miksch, H. Schumann, and C. Tominski, "Visualization of time-oriented data," *Human-Computer Interaction Series*, London : Springer-Verlag, 2011, pp.45-68.

[17] P. Peltoniemi, "Is it possible to study scientific concepts? ," in *Proc. the 8th International Conference on Terminology and Knowledge Engineering* , 2008, pp.123-136.

[18] U. Eco, *The infinity of lists*. London: Maclehose, 2009.

[19] J. M. Bocheński, *The Methods of Contemporary Thoug*ht, New York: Harper Torchbooks, 1968, pp. 84-86.

[20] S. Bertschi, S. Bresciani, T. Crawford, R. Goebel, W. Kienreich, M. Lindner, V. Sabol, and A. V. Moere, "What is knowledge visualization? perspectives on an emerging discipline," in *Proc. the 15th Internationa*l *Conference on Information Visualisation*, 2011, pp. 329 – 336

[21] W. Kienreich. (July 2006). Information and knowledge visualisation: An oblique view. Available: http://www.map.archi.fr/mia/journal/articles/vol0/num1/kienreich.pdf

[22] R. Spence, *Information Visualization*, New York: Addison-Wesley, 2001.

[23] C. Ware and R. Bobrow, "Supporting visual queries on medium-sized node-link diagrams," *Journal Information Visualization*, vol. 4, no. 1, pp. 49-58, Spring 2005.

[24] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Proc. the Third*

International AAAI Conference on Weblogs and Social Media, 2009, pp. 361-362.
[25] E. R. Tufte, *The Visual Display of Quantitative Information*, Cheshire: Graphic Press, 2001.
[26] J. Maeda, *The Laws of Simplicity*, Cambridge: MIT Press, 2006.
[27] G. Schuller, *Designing Universal Knowledge*, Baden: Lars Müller Publishers, 2009.
[28] J.Y. Blaise, I. Dudek. "Can simplicity help? ," in *Proc. the 14th International Conference on Knowledge Technologies and Data-driven Busines*, 2014.

Iwona Dudek is with architect (Cracow's Technical University), a PhD history of architecture and urbanism (Faculty of Architecture, Cracow's Technical University, 2001) - since 2001 is a CNRS senior researcher at CNRS and contributes to the UMR MAP 3495 CNRS/MCC research team in Marseilles.

Her research focuses on InfoVis as applied to historical sciences, and particularly to historic architecture. Her research themes and involvements include knowledge modelling, history of architecture and urban forms, diachronic analysis of historic architecture, visual analytics as applied to historical sciences, time-oriented data/information management, spatio-temporal information systems, graphic semiology in the context of InfoVis techniques and tangible interfaces.

Dr. Dudek is the author and co-author of over 70 publications on topics ranging from history of architecture to information management and InfoVis. She acts as a reviewer or international scientific journals and conferences.

Jean-Yves Blaise, Architect ENSAIS (École Nationale Supérieure des Arts et Industries de Strasbourg), PhD Mathematics & Informatics (Paul Cézanne University Aix-Marseille III, 2003) - since 2005 is a senior researcher for CNRS (French National Body for Scientific Research).

His research focus on knowledge representation and infovis applied to historical sciences, and particularly to historic architecture. Research themes and involvements include ontology and terminology for architectural concepts, spatio-temporal information systems, visual analytics as applied to historical sciences, time-oriented data, architectural morphology analysis, and tangible interfaces (tactichronie patent).

Dr Blaise wrote and published over 75 scientific papers bridging disciplinary gaps in a wide range of scientific disciplines (KR, data analytics, computer graphics, Infovis, digital heritage, Humanities, *etc*.). Acts as a reviewer in the above-mentioned scientific circles.