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

Iwona DUDEK , Jean-Yves BLAISE
UMR 3495 CNRS/MC MAP – Marseille, France

Abstract

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

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

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

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

A**ppendix 1**

groups of activities vs. data lifecycle

The division of activities into the five groups is sometimes perceived – especially in data science circles - as an approach based on data lifecycle[1].It would be dishonest not to deny this fact: our work was in no way inspired by this ambition. However, a close look at approaches in data lifecycle management allowed us to draw some interesting conclusions and to learn something new.

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

The practices focused on data lifecycles or persistence are conceptualised differently in different communities. Although they share similar objectives (*i.e.* to manage data throughout its existence), they vary in terms of the granularity of the problem presentation (containing a different number of stages), the type of data involved (*e.g.,* business, government related data, scientific data) and the approaches (type of actions). Even a cursory analysis of a selection 'data lifecycle' diagrams (Fig 5) reveals the absence of consistent regularities of the succession of phases proposed in different contexts and for different data – only the position of a 'data acquisition steps' (red colour) might be seen as an example of a common trend. The presence of a deletion phase present only in some procedures is also revealing. In some data management groups, the issue of data overflow and thus the need for regular data filtering and scheduled deletion is already noted and integrated – in some communities only.

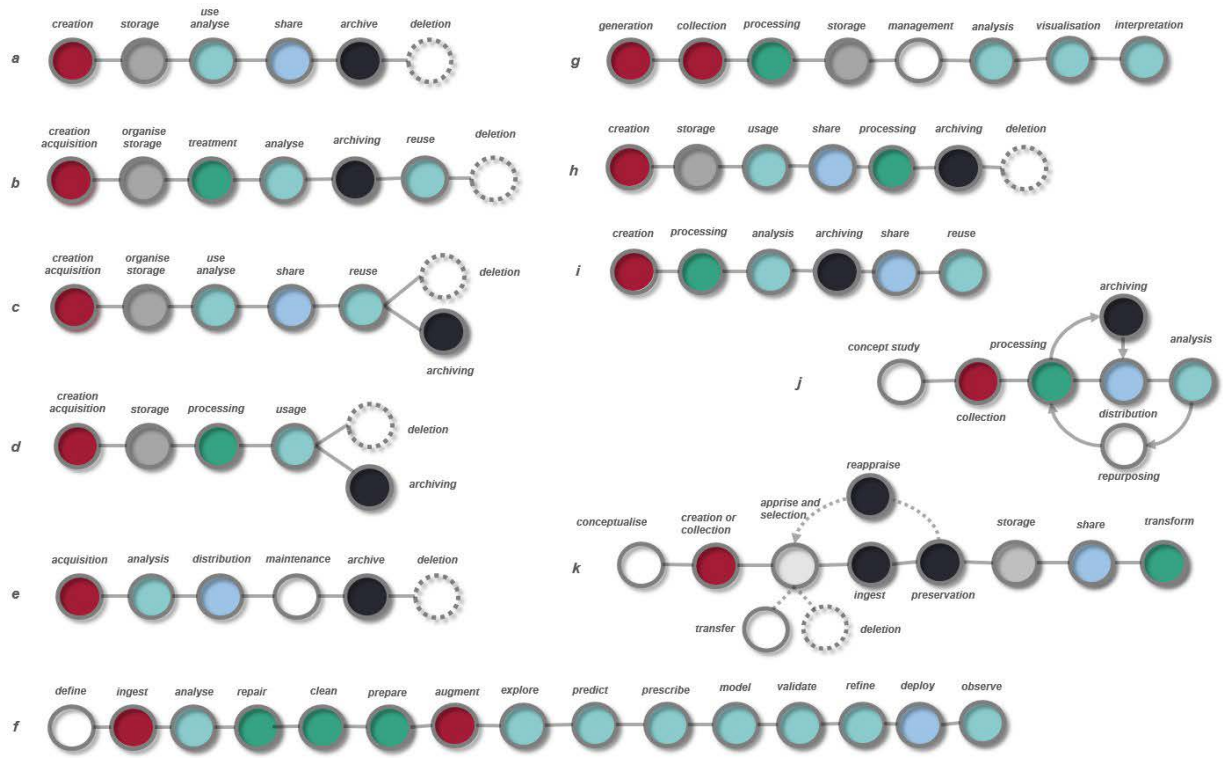
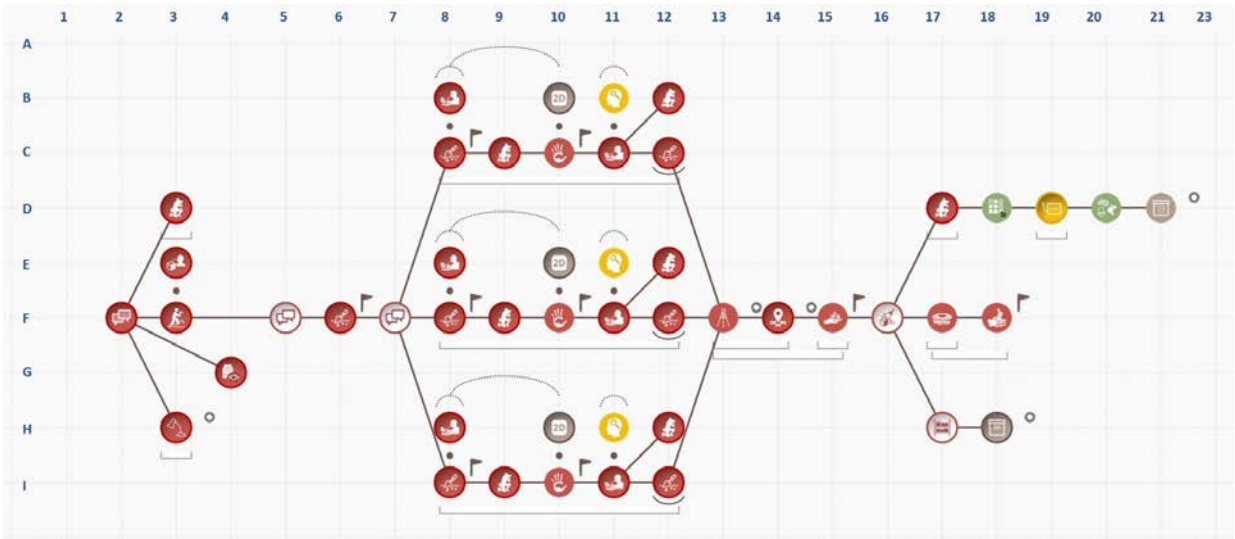


Figure a1 *– Eleven graphs representing successive phases in data flow management (progression from left to right) put forward for: business data management (a- IBM, b- talented.com, d - startupgeek, g- Harvard Business School), geospatial data management (e- Sanborn), governmental data policy (c- NSW government), data management and analytics (h- ClicData, f- Trust Insights), research data (i- Nanyang Technical University), data curation (k- Digital Curation Centre), medical data (j- University of Nebraska). Original graphics have been redrawn and interpreted by authors – colours were introduced to underline order of steps dedicated to similar goal-oriented activities (e.g., the 'ingest' phase in schemes k and f is denoted by a different colour, as the same term is used in these cases in other contexts. It is difficult to discern consistent regularities of the series of actions proposed in different contexts and for different data. The presence of a deletion phase present only in some procedures is also a tell-tale sign.*

Appendix 2

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



The initial phase of the operation begins with a discussion between the participants (*conjecture*) (**F2**).

From this point onwards, four parallel activities are undertaken: (**D3**) a photographic survey of the chapel, the facades of the houses in the hamlet, the site, ... - a repetitive activity before and during the excavation, (**F3**) surface exploration during the preliminary visit combined with (**E3**) potential on-site observation, (**G4**) non-intrusive exploration of the chapel - possibly before and during excavation, (**H3**) initiation of the documentary research - possibly before and during excavation.

(**F5**) possible organisational discussions after the pre-visit, followed by (**F6**) outset of excavation – opening of three parallel trenches (mechanical shovel). From now on, excavation work is carried out in parallel (*conjecture*) in three trenches.

Work carried out in each of the trench concerned (*hypothesis*): (**C-F-I8**) iterative hand excavation combined with (**B-E-H8**) subsurface observation, followed by (**C-F-I9**) photographic documentation and (**C-F-I10**/**B-E-H10**) inventory of stratigraphic units (measurements and documentation). The whole sequence is repeated until an artefact is discovered. In this case finds are examined (*first interpretations?*) (**C-F-I11**/**B-E-H11**) photographed (**B-E-H12**). Than follows a cautious hand excavation process (**C-F-I12**).

In the case of finds collecting, the following activities are carried out (*hypothesis*): (**F13**) a topographic survey of the test pits, (**F14**) referencing localisation (of each artefact), (**F15**) and finds collection (repetitive sequences of activities) - the finds are then possibly cleaned (**F16**),

From this point, three parallel activities are undertaken (*conjecture*): (**F17**) site cover up and backfilling (**F18**) - carried out successively trench by trench (*conjecture*), (**D17**) photographic documentation of closing operations, followed by possible selection of photographs for storage (**D18**), their annotation (**D19**), classification of photographs taken during the operation (**D20**) and digital archiving of the entire collection (**D21**), as well as on-site storage of finds (**H17**) and physical archiving (**H18**).