Research workflows and digital data sets: traceability, reproducibility, comparability issues illustrated on spatio-acoustic data acquisition protocols

Iwona Dudek

Today, I'd like to share with you some ideas and experiences in describing research workflows and digital data sets, with a focus on traceability, reproducibility and comparability issues. I will illustrate my points with spatio-acoustic data acquisition protocols.

When we try to explain how we managed to achieve a result (e.g., production of a given data set) it implies the choice of language to be used. Description effort firstly involves the choice of a language – quite often it is an ethnic language such as English, French or Italian.
Ethnic (or natural) languages fall somewhere between '*hard languages*' (context-free languages where the level of poly-interpretation is absent or low) and 'soft languages', that are strongly poly-interpretative (where meaning is highly subjective and context-dependent).
If we want to ensure reproducibility and comparability, we need to reduce the possibilities of poly-interpretation of their description.
This can be done by the means of conceptual modelling, controlled vocabulary and diagrammatic representations. This is the basis on which the MEMORIA exploratory information system is built.

The structure of my presentation can be summarised in five points. I will start with the brief outlain of our research objectives. Then, I will present you some underling concepts that will be illustrated by the means of examples. Finally, I will present some of the advantages and limits of the system.

MEMORIA is an exploratory web-based information system built to address the burning issue of traceability, verifiability, comparability and potential reproducibility of research results and protocols. It aims to provide a practical solution for formalising and describing research workflows.
What are the main objectives?
The system should allow us to identify, structure and preserve information about different types of research results with their metadata and to identify, structure and preserve information about the sequences of actions that led to their creation.Such a workflow can be represented by purely verbal means. We have decided to depict it using a diagrammatic representation combining paradata. We will call it ***memoria workflow diagram***.
This diagram is a sort of graph where the nodes represent activities and connections between them represent time-oriented relations between activities. Activities may be organised as a chain. Some sequences of activities may run concurrently in time. Activities may be combined – it means they take place simultaneously or they occur without any apparent sequential order. Iteration is another type of temporal relation. And it differs from a repetition. Each node of a diagram represents an activity. Its colour refers to one of five groups of activities. Within each group, activities are structured, defined and exemplified. They result from successive knowledge elicitation steps - it means that the activities structure can be enriched.
Activities are organised in a workflow diagram using a drag-and-drop web interface, as illustrated here. The first step is identification and selection of activities. Then the activities are positioned inside the grid. Finally we define the relations between activities and we identify the moments where the results appear. The next step is a description of individual activities (activity by activity) by:

- identifying particularities of this activity (controlled vocabulary),
- as well as free remarks and comments

Once an activity is characterised, the background of its icon appears.

A workflow diagram structures and organises sequences of activities together with their paradata. It is a sort of a backbone for the 'body of the process'.
Paradata organised by the mean of a workflow are stored in DB. They are freely accessible, searchable, and editable. They can be queried, compared and visualised.

Here we have two workflow diagrams representing acquisition protocols: acoustic and spatio-acoustic data acquisition. We easily can notice differences of their complexity, their structure, groups of activities involved within each process and so on.
The difference is even more striking if we compare processes behind results of different nature.

To exemplify the approach I will focus on this spatio-acoustic data acquisition protocol (video showing the protol recorded in situ).
What are all those people doing? Who works with whom? It is in fact difficult to say, because different people are doing different things. Maybe it is hard to believe, but they do it in an organised way. However, some actions are planned and others spontaneous.
What you see it is one interdisciplinary acquisition protocol made of numerous actions.
This actions may be expressed in a form of memoria workflow diagram, that can be used to fix the order of sequences of activities corresponding to:

- instruments positioning,
- documentation,
- acoustic data acquisition and pre-processing,
- metric and visual data acquisition.

Such a diagram allows both: context view analysis and the focus on particularities of each activity.
The same data can be visualised in different manners. On the left process represented as a workflow diagram. On the right, the same process is depicted as *activities' proportion ring* – the visualisation that displays:

- what are the relative proportions of each group of activities inside a process,
- the links to the results and inputs of this process,
- information about the instruments and softwares that have been used
- ...

The same data within a process variability chord diagram visualisation are exploited to display relations between processes conducted within a given project.
Project A, project B project C.

Let's pass to the conclusions.

- This documentation framework enables: tractability of results, verifiability of research protocols, comparability and analysis of individual processes. Reproducibility of results is another issue, as it does not depend only on the quality of the workflow description. Some results are simply not reproducible.
- The system may be used to support of the planning of research protocols based on past experiences (whether they are interdisciplinary or not).

- It may be used as a theoretical and/or tacit knowledge-sharing environment targeting educational or research communities.
- It may be further developed and exploited as visual analytical framework.
- Finally, by requiring careful thinking, it can help us to work more efficiently, more consciously and in a less routine way.

They are to groups of constrains and limits of the system: acceptability constraints and the technical constrains.

- MEMORIA remains a human-centred system, we should not forget that quality and reliability of the visualisations depend largely on the data entered by various individuals.
- Documenting is always time-consuming. ... in this case we need to 'learn the language' of this system and gain some confidence in using it. The learning curve for some than for others.

The final point concerns our willingness to share information with others and the extent to which we are prepared to do so.
The technical constrains relates to sustainability of web-based information systems and costly and complex development steps.

We have some future works in mind which I'll be happy to discuss with you later today. The number of issues that can be covered is not limited to cultural heritage. It may involve the experimental sciences, applied sciences, information technology, the humanities and the social sciences.

Abstract
In this presentation, we discuss the importance of traceability, verifiability and comparability of research workflows and digital data sets. We highlight the challenges associated with the choice of language, which can affect the interpretation of research descriptions. To address this, we propose the use of conceptual modelling, controlled vocabulary and diagrammatic representations as a means of reducing interpretation variability. These elements form the basis on which the MEMORIA exploratory information system is built.
The main objectives of the system, notably the identification and preservation of information on different types of research results and the sequences of actions that led to their creation, are illustrated using spatio-acoustic data acquisition protocols.
We conclude by mentioning the constraints and limitations of the system, including the reliability of the data and the time-consuming nature of the documentation.

Résume
Dans cette présentation, nous discutons de l'importance de la traçabilité, de la vérifiabilité et de la comparabilité dans les processus de recherche et les ensembles de données numériques. Nous soulignons les défis liés au choix de la langue, qui peut affecter l'interprétation des descriptions de recherche. Pour y remédier, nous proposons d'utiliser la modélisation conceptuelle, le vocabulaire contrôlé et les représentations diagrammatiques comme moyen de réduire la variabilité de l'interprétation. Ces éléments constituent la base sur laquelle est construit le système d'information exploratoire MEMORIA visant à formaliser et à décrire les flux de travail de la recherche.
Les principaux objectifs du système, notamment l'identification et la conservation d'informations sur différents types de résultats de recherche et les séquences d'actions qui ont conduit à leur création,

seront illustrés sur des diagrammes de flux de travail pour les protocoles d'acquisition de données acoustiques et spatio-acoustiques et diverses manières de visualiser les données collectées.
Dans la conclusion, nous mentionnons les contraintes et les limites du système, y compris la fiabilité des données et la nature chronophage de la documentation.

Iwona Dudek is a research fellow at the MAP CNRS research team in Marseille. Her research focuses on the application of information visualisation (InfoVis) to the historical sciences, in particular historical architecture. Her research interests cover a range of topics, including data acquisition, knowledge modelling, the history of architecture and urban forms, visual analysis, time-oriented data management, spatio-temporal information systems and graphic semiology in the context of InfoVis techniques and tangible interfaces. She has authored and co-authored over 70 peer-reviewed publications on topics ranging from architectural history to information management and InfoVis.
Since 2014, she is focusing also on epistemological issues, deontology, ethics and scientific integrity issues.