# An Ontology for Supporting Digital Forensics Controlled Experiments: Early Stages of Development

**Thiago J. Silva**[1], **Edson OliveiraJr**[1]

[1]Informatics Department – State University of Maringá (UEM), Maringá – PR – Brazil

josthiago1@gmail.com, edson@din.uem.br

***Abstract.*** *The experimentation process is one of the main means of science to evaluate theories based on hypothesis. Science evolves most taking into account the performing of controlled experiments, thus providing trust evidence for different research fields. However, for the Digital Forensics (DF) field, formal controlled experimentation has been neglected over the years. In a recent systematic mapping of the literature, we found more than 200 experiments with few formalization of their procedures, thus jeopardizing its evidence reliability and the capacity of reproducibility. Therefore, this paper provides early steps to specify an ontology for supporting proper planning, conducting, and dissemination of DF controlled experiments. The ontology has as a basis a conceptual model created to specify the main elements of an experiment. We adopted the Uschold and King approach to develop such an ontology. We first designed the ontology with the protégé ontology editor, as it organizes knowledge. In addition, the WebVOWL tool was also used from the json file generated in the protegé tool, as this aids at the construction of a dynamic visual identity for the ontology. As general results, we understand that despite the number of ontologies in the literature being relevant, few present a structure that satisfies a relation of similar objects for their properties, in addition do not covering the context of a DF experiment. Therefore, despite the ontology construction is in its early stages, it is expected that it will shad light to the field of experimentation in DF throughout a hierarchy and formalized process.*

## 1. Introduction

Controlled experimentation is a reliable scientific method based on empirical assumptions to define and evaluate a certain theory and/or its practical purpose. It is performed by running hypotheses testing techniques over well-defined variables and procedures [Tedre and Moisseinen 2014]. In addition, proper documentation is essential to promote its reproducibility [OliveiraJr et al. 2020].

For the Digital Forensics (DF) area, controlled experimentation is straightforward needed as DF is currently a world-widely trend research area. However, it has been noticed experimentation for this area is lacking dedicated attention to its proper application, thus providing non-reliable evidence [Casey 2013, OliveiraJr et al. 2022]. By proper application we mean, for instance, a well-defined variable set definition, randomness, accurate definition of objective, choose of descriptive and inferential statistical techniques, extreme values and outliers analysis, and packaging and dissemination procedures.

In general, DF experiments provide poor analysis of the collected data mostly based on the mean of the samples, which is not sufficient for providing a proper and trusty

result analysis and discussion [OliveiraJr et al. 2022, OliveiraJr et al. 2021, Casey 2013]. Therefore, we hypothesize that the more formal is the experiment, the more reliable and reproducible are its results. By formal we understand an experiment needs at least to be well-documented taking into account all experimental elements of its protocol, making it consistent for result analysis and discussion, thus dissemination.

Our DF research group has already developed and evaluated a conceptual model (see Section 2) to supporting defining, conducting, and disseminating DF controlled experiments [OliveiraJr et al. 2020]. Although a conceptual model might aid DF researchers and practitioners to conduct an experiment, we find it is missing a formal support for data and metadata levels of the conceptual model. In addition, there are also limitations regarding the performance of crimes and ethical principles, which are relevant, thus the results do not conflict with formal investigations and do not harm privacy principles. Therefore, we understand an ontology might be a way for providing such a formalization, as it is a means to support the construction of knowledge, or even the semantic classification of the conceptual model concepts.

Based on a recent Systematic Literature Review (SLR) we performed[1], the literature does not present any DF controlled experimentation support ontology. An ontology use example might be the lack of experimentation in Great Britain in around 90% of modern crime cases [Overill and Collie 2020], thus making them error-prone due to investigations without standardization and tool failures [Alvarez 2011]. Our ontology might help, for instance, to prevent judgments from being influenced by hasty or incorrect experimental forensic conclusions. Therefore, such an ontology aids to organize experimental data towards providing a proper documentation of controlled experiments and effective report based on a reliable methodology and domain.

This paper aims at introducing the design of an ontology for DF experiments formalization. The ontology development is centered on the [Uschold and King 1995] method, which focuses on four essential phases for the ontology development: identification, construction, evaluation, and documentation.

## 2. Experimentation in Digital Forensics

Experimentation can be seen as a straightforward way to provide reliable and reproducible process and evidence [Tedre and Moisseinen 2014].

An experiment is set based on five well-defined phases [Wohlin et al. 2012]: **Definition**: should describe the experiment objective, purpose, focus, perspective, and context; **Planning**: defines the experiment protocol to be carried out. To do so, it describes: hypotheses, independent and dependent variables, selection of participants, experimental design based on variables, and threats to validity; **Operation**:it consists of three processes: data preparation, data collection, and data validation; **Analysis & Interpretation**: data is analyzed based on descriptive and inferential statistics, thus providing interpretation based on the hypotheses set; **Presentation & Package**: it is the documentation and presentation of the results. It is at this point that the learning acquired during the execution of the experiment must be clearly reported.

Although there is a tremendous effort in the general science experimentation

[1]Currently under review in a journal.

process, for the DF research and practice areas, experimentation has been neglected [Casey 2013, OliveiraJr et al. 2022], thus needing for a more formal planning and conduction.

To provide a way to promote experimentation in DF, [OliveiraJr et al. 2020] proposed a conceptual model for DF experiments. Such model seeks to demonstrate its concepts based on the experimentation process, regarding six main elements: `DF Experiment`; `Planning`; `Pre-Operation`; `Operation`; `Analysis and Interpretation`, and `Dissemination`. Figure 1[2] depicts the overall model view[3].

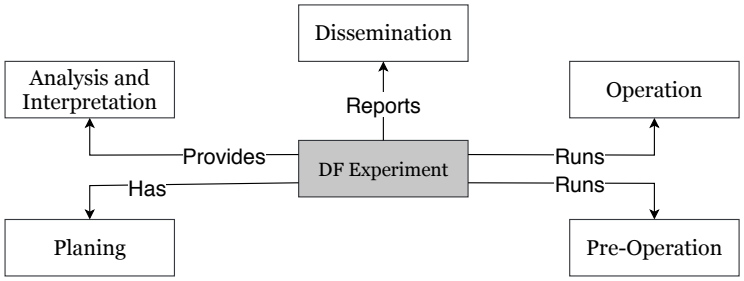


**Figure 1. Conceptual Model for Supporting DF Controlled Experimentation - Overall View [OliveiraJr et al. 2020]**

**Planning** refers to the preparation of the experiment and, therefore, is related to the main details and stages of building a scenario for the case to be analyzed. It is related to the concepts of hypotheses, participants, analysis, reporting, acquisition, examination phases, replication of experiments, type of experiment, experimental unit, variables, type of project, instrument, and objective.

**Pre-Operation** defines the setup of the experiment, as well as hardware, algorithm, and software. In hardware it focuses on volatile or persistent memory components (virtual or physical). Algorithms, on the other hand, are parameter-dependent and are means of analysis with well-defined principles. Software is related to a virtual environment, operation or application of the experiment. At the top of this phase we have performance issues such as: test training; pilot project, which performs the initial experiment and assesses the possible inconsistencies that the execution may present; and Benchmark to aid in the planning process.

**Operation** This phase is related to the organization and execution of experimental activities. The intention is to collect original or duplicate data in an appropriate way, having materials that can lead this process. In addition, it is at this moment the experimenter and the participant are introduced.

The **Analysis and Interpretation** phase has key concepts that are the basis for developing and understanding the data. For that, there is the technical analysis and plotting of data, limitation, evaluation and validity of the experiment. Data plotting is a performance feature and seeks to express values visually, using a table, bar graph and others. And as a way of evaluating the values, quantitative and qualitative analyzes were defined. When it comes to experiments, it is also important to highlight the limitations and threats

[2]All high-resolution figures are at `https://doi.org/10.5281/zenodo.6396707`

[3]Conceptual model complete view is at `https://doi.org/10.5281/zenodo.6341020`

to validity discussed in the analysis and interpretation of data. These procedures are necessary for an important statistical result about the results.

**Dissemination** considers actions such as diary/note, experimental issues and data forensics management plan. Within the dissemination process there is also a dataset composed of a unique identification, authorship, citation and a repository for storage of the experiment and data, whether these are quantitative or qualitative. This is important to maintain data integrity and have a reliable means of easy access.

This model might be used: (i) for primarily guiding one to plan, conduct, and disseminate an experiment; (ii) as a checklist for specific elements of one of the five concepts; (iii) for re-engineering existing experiments and improve them towards making them reproducible; and (iv) for developing tool support for DF experimentation.

## 3. Towards an Ontology for DF Experiments

The main idea is to create an ontology to support DF experimentation formalization based on the conceptual model of Section 2. It is because such model covers the DF phases of identification, preservation, collection, examination, analysis, and presentation of DF evidence.

To do so, we followed the ontology creation process by [Uschold and King 1995]. We firstly mapped the high-level concepts of the conceptual model to elements of the ontology. We performed a cascade analysis of the conceptual model data and evaluated the requirements and integration of them to incorporate into the ontology. Secondly, we modeled the main phases of the DF process to first-class elements of the ontology.

For modeling the high-level view of the ontology (Figure 2), we constrained the main classes: `Analysis and interpretation`, `dissemination`, and `planning`. We made this as these phases bring greater conciseness in the integration of the ontology elements.

Our ontology is being developed using the Protégé[4] tool with a general model composed of seven phases as follows:

- **Experiment:** refers to the controlled experiment to be carried out. It is in the center of the model as we can infer on next steps of the DF process;
- **Identification:** it aims to evaluate the incident, in addition to analyzing whether it has any relationship with other incident already analyzed;
- **Acquisition:** it is the process of retaining the data storage locations of the experiment;
- **Preparation:** the environments are elaborated and configured, thus the experiment can be executed;
- **Preservation:** it is responsible for guarding the evidence;
- **Analysis and Reconstruction:** it encompasses experimental analysis techniques;
- **Review and presentation:** after completing all phases of the experiment, the objective is that it can be reviewed and disseminated through reports and presentation.

[4]`https://protege.stanford.edu`

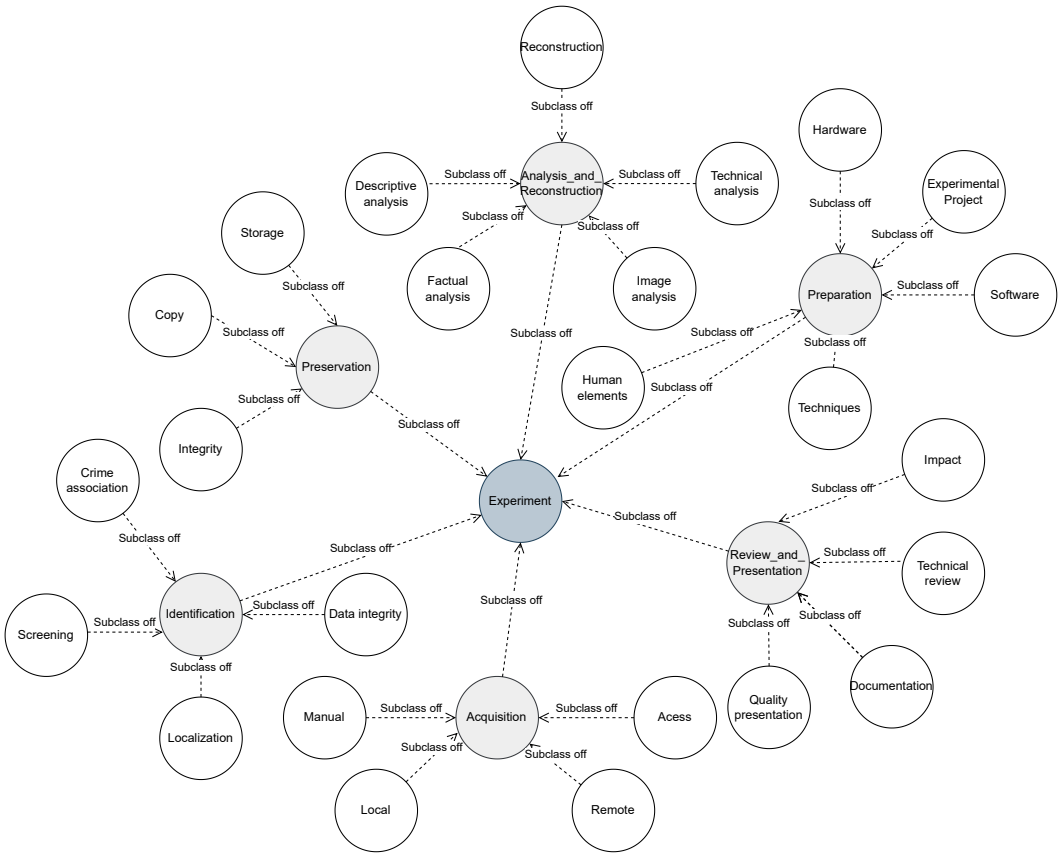


**Figure 2. The Ontology High-level View**

We expanded this high-level model (Figure 2) with subclasses, which are still under development. Our intention is to provide low-level elements, thus a DF experiment might be well documented at the level of subclasses and properties.

We, then, created what we so-called a low-level view of the ontology (Figure 3) based on OWL 2 language. With such a view, we present potential classes, subclasses, and their relationships.

In general, a DF experiment is centered on the ontology and highlighted by a darker color (gray), while its subclasses are identified by the light gray tone. As for the circles that are with the symbol "n" and highlighted in blue, these refer to the properties of an object as it makes the relationship of an individual with another similar. This also occurs with the property highlighted in the screening subphase. However this is not a property of the object, but of the individual, since it can have a data type in byte or string.

For the development of the low-level ontology view, we used two source files: one from [OliveiraJr et al. 2021] at `http://doi.org/10.5281/zenodo.3929921` and one from [OliveiraJr et al. 2022] at `http://doi.org/10.5281/zenodo.3835139`. In the first two works, more than 200 DF experiments were analyzed aiming at identifying their rigorousness at following the experimentation process, protocols, and dissemination policies.

An example of how to "read" the ontology is as follows based on the **Review and Presentation** DF phase, represented by the class `Review_and_Presentation`. In

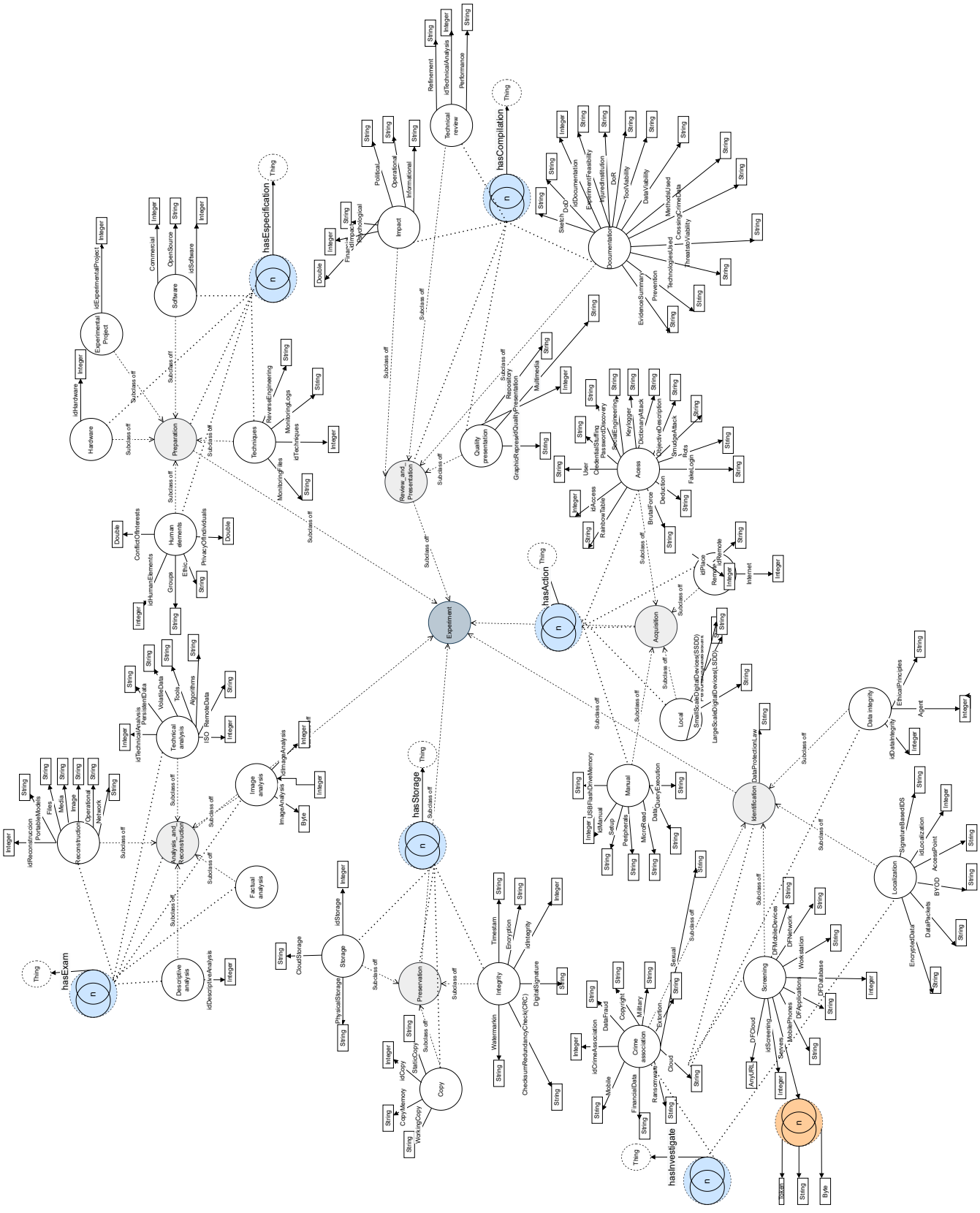

**Figure 3. Low-level View of the Ontology**

this phase, the experiment can be examined and exposed to the public. At this moment, the experiment dissemination is validated, analyzing the elements used and its hypothesis. This phase is composed of the sub-phases (right-lower side of Figure 3) `Quality presentation`, `Documentation`, `Technical review` and `Impact`.

The `Impact` sub-phase refers to the consequences a cybercrime. This might have properties as: *Informational*, which considered a discovered knowledge and the dissemination of information; *Operational* consistent with the procedures to be performed in the experiment, qualifying data to obtain results; *Political*, which is an ideological influence that brings individual motivation to the criminal act; *Psychological* as a reputational damage or means of belittling the exposed event; and *Financial*, economic aspects to minimize the profitability of the company or individual.

`Documentation` is essential after the conclusion of the experiments. They are properly recorded by means of properties such as: *Prevention*, which is a description of which possible points have been identified and can be improved through measures; *TechnologiesUsed*, refers to the digital media incorporated during the experiment,; *ThreatstoViability*, refers to possible issues that may hamper the effectiveness of the experiment; and *MethodsUsed*, methods employed in the experiment to plot data such as histogram and bar graph.

`Quality presentation` is how the quality data of the experimental process is presented as, for instance, *GraphicRepresentations*, which are visual ways of sketching the data of the experiment; *Multimedia*, programs and systems that can sketch in a practical way some important point of the ontology; and *Repository*, which refers to a trusted location.

The `Technical review` phase evaluates the results that are complemented by: a *Refinement*, which is a strategic way of adjusting the list of items to estimate the time needed for the results and their processes to be evaluated; and *Performance*, which seeks to understand whether the compilation of the experiment presents a waste of time and vulnerability for scalability [Villar-Vega et al. 2019].

As we can see, the ontology is not yet finished, as several elements from the conceptual model and from the DF domain are missing representation. Therefore, we discuss the next steps of this work in the next section.

## 4. Next Steps for the Ontology Development

We envision certain steps to finish the development of our proposed ontology as follows.

Model the DF sub-phases that are not yet present in the ontology, such as, hardware/software and descriptive analysis. For this, a new documentary analysis of a qualitative nature will be carried out in the extraction files originally used to starting modeling our ontology. In addition, we are working on the ontology review, with the aid of experts, thus we can assess whether it is possible to include other phases and better plan, operate and disseminate DF experimentation.

We are currently using the Protégé tool to validate the ontology and creating scripts to populate the ontology, as well as to query and infer on the ontology. With such scripts, one might be able to properly use the ontology to search for DF experiments and to store experimental data of existing r prospective experiments.

An empirical evaluation is currently been performed in our first version of the ontology, in the form of a qualitative study based on grounded theory, thus aiming at pointing out important aspects to motivate the use of the ontology, as well as missing elements.

As a last step, we intend to run a controlled experiment with DF researchers who authored any DF experiment in the literature aiming at analyzing the feasibility of the ontology based on the Technology Acceptance Model (TAM). TAM provides a way to assess the level of perceived usefulness and perceived ease-of-use. However, the experiment will still be carried out, since the ontology is in the construction and evaluation phase, with no tangible results that can be presented for now.

## 5. Final Remarks

We presented two ontology views: a high-level one with the seven main elements; and a low-level one with the expanded elements from the first view. These two views have been giving us the opportunity to analyze how we can contribute to the DF experimentation area based on an initial novel conceptual model recently developed. We also presented the next steps to fulfill the development of our ontology, most based on empirical activities. Experts are already playing an important role at the development of the ontology as they are currently analyzing it.

## References


Alvarez, L. (2011). Software designer reports error in anthony trial. *New York Times*, 19.

Casey, E. (2013). Experimental design challenges in digital forensics. *Digital Investigation*, 9(3):167–169.

OliveiraJr, E., Silva, T., Zorzo, A., and Neu, C. (2022). Digital forensics experimentation: Analysis and recommendations. *Forensic science review*, 34(1):21–41.

OliveiraJr, E., Zorzo, A. F., and Neu, C. V. (2020). Towards a conceptual model for promoting digital forensics experiments. *Forensic Science International: Digital Investigation*, 35:301014.

OliveiraJr, E., Zorzo, A. F., and Neu, C. V. (2021). Experimentation of digital multimedia forensics: State of the art and research gaps. *Wiley Interdisciplinary Reviews: Forensic Science*, page e1406.

Overill, R. and Collie, J. (2020). Deep: Extending the digital forensics process model for criminal investigations. *Athens Journal of Sciences*, 7(4):225–240.

Tedre, M. and Moisseinen, N. (2014). Experiments in computing: A survey. *The Scientific World Journal*, 2014(1):1–12.

Uschold, M. and King, M. (1995). Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing*, pages 1–15.

Villar-Vega, H., Perez-Lopez, L., and Moreno-Sanchez, J. (2019). Computer forensic analysis protocols review focused on digital evidence recovery in hard disks devices. In *Journal of Physics: Conference Series*, volume 1418, page 012008. IOP Publishing.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.