Towards a human-in-the-loop curation: A qualitative perspective

Alejandro Adorjan, Genoveva Vargas-Solar, Regina Motz

Towards a human-in-the-loop curation: A qualitative perspective

Alejandro Adorjan
Facultad de Ingeniería
Universidad ORT Uruguay
Montevideo, Uruguay
adorjan@ort.edu.uy

Genoveva Vargas-Solar
CNRS, Univ Lyon, INSA Lyon, UCBL,
LIRIS, UMR5205
Lyon, France
genoveva.vargas-solar@cnrs.fr

Regina Motz
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
rmotz@fing.edu.uy

***Abstract*—This paper proposes a data curation environment to record, maintain, and enrich research data using quantitative and qualitative methodologies. The research addresses the following research questions: What is research data curation from a hybrid perspective? and What software tools are adapted for hybrid research projects? The paper also proposes a hybrid research workflow representing the phases of projects adopting this kind of methodology towards a human-in-the-loop approach. It introduces a view model to represent the data produced across its stages, which should be curated. Finally, proposes a set of operators to manage and explore the different versions of curated data and their associated knowledge.**

***Index Terms*—Hybrid research, data curation, semantic enrichment**

I. Introduction

Data-driven experiments devoted to studying social and humanities problems can potentially benefit from using data science techniques, including software, algorithmic and data management tools. Social and humanities studies address research problems by combining qualitative and quantitative methods. Qualitative research seeks to understand a given research problem or topic from the local population's perspective and systematically uses a predefined set of procedures to answer the question. Therefore, it collects evidence and produces findings not determined in advance and applicable beyond the study's immediate boundaries. Qualitative research effectively obtains culturally specific information about particular populations' values, opinions, behaviours, and social contexts. It also effectively identifies intangible factors, such as social norms, socioeconomic status, gender roles, ethnicity, and religion, whose role in the research issue may not be readily apparent. When used along with quantitative methods, qualitative research can help to interpret and better understand the complex reality of a given situation and the implications of quantitative data.

Consider the following example: a study willing to analyse street posters and graffiti to observe/exhibit the moral value system hidden or covered up in this artistic experience [1]. The research question that guides the study of street posters and graffiti considers several aspects. First, it explores whether it is possible to find textual remains and use other codes that have been recorded from the seminal text. Next, it analyses whether these codes now dialogue in a spatio-temporal relationship as a narrative of different elements and figures introduced in an incomplete form as quotations, references, and styles that the reader will later take as homogeneous.

The research project includes organising a data harvesting campaign that implies choosing the urban spaces to visit to chase pieces to analyse the pictorial narratives. A methodological strategy to harvest pictures means defining inclusion and exclusion criteria, angles and techniques to take photographs, spatio-temporal tagging of the photos and comments on the choices and reasons why researcher considers that a particular piece fulfils inclusion criteria. Furthermore, the research team can also organise interviews with civilians and art specialists to harvest their interpretations and comments about the pictures. The harvested pictures tagged with qualitative and quantitative metadata can then be classified manually or automatically. The content can be further processed to compare colours, salient objects, and spatial distribution of salient objects to try to identify narratives and correlate them with seminal texts connected with the pictures' content and the comments from interviews.

According to observations and findings, the research team can agree to adjust the initial research questions, harvesting criteria, interview content and target people. The criteria for making decisions and adjustments may also be gathered as research data. New harvested pictures, observations and interpretations can then be compared with previous findings. These actions can be repeated until the research team converges on an acceptable method with criteria and adapted tools (e.g., interviews, choice of data producers) to harvest data, observations, findings, plots and metaphors that can lead to reporting qualitative and quantitative results.

Research projects that combine qualitative and quantitative methods must keep track of the process that leads to final results, the decision-making milestones and criteria, and adjustments. This information exhibits the method that supports the validity of final results and interpretations. Keeping track of this information is also part of the knowledge generated through the project, and it can promote reproducibility or serve as guidelines for projects with similar characteristics.

Supported by the National Agency for Research and Innovation (ANII), Uruguay.

Data curation is critical for this kind of research project where human participation is omnipresent at all stages of the research workflow [2]. Managing research data is a core responsibility of data curation, a sub-discipline within the library and information sciences. Generally, research projects do not explicitly design and model curation guidelines to keep track of a research project's evolution, which can help replicate an experiment. The absence of curation methods, practices, techniques, or standardised frameworks adapted for hybrid methods, including qualitative and quantitative strategies, challenges a researcher to manage data collections efficiently. Therefore, a new research opportunity arises in data science to provide methods, guidelines, practices, and applications that facilitate exploring and curating qualitative data.

Our research aims to develop a data curation environment to record, maintain, and enrich research data using hybrid, quantitative and qualitative methodologies. The research addresses the following research questions:
RQ_1: What is research data curation from a hybrid perspective? and RQ_2: What software tools are adapted for hybridresearch projects?
This paper introduces our approach to providing a human-in-the-loop curation environment adapted for hybrid (qualitative and quantitative) research projects. Accordingly, the paper is organised as follows. Section II introduces the background and related work. Section III introduces the curation environment for hybrid research projects data with human-in-the-loop (HITL) that we propose. Section IV shows how to curate hybrid research data and knowledge using our approach in a use case. Finally, Section V concludes the paper and enumerates future work.

II. Background

This section characterises and compares qualitative and quantitative research methods that are the background of our work. Then, presents the principle of data curation, describing the kind of tasks associated with it when applied to scientific experiments. Finally, the section introduces related work regarding Qualitative Data Analysis Software (QDAS), the scope of solutions and their limitations for semi-automating hybrid research projects.

A. Data-driven qualitative and quantitative research methods

Quantitative data-driven methods study phenomena or objects of study based on observations collected a priori and promote a deductive approach for driving conclusions by processing data using numerical methods, statistics, machine learning and other artificial intelligence models.

Qualitative approaches promote inductive methods that interleave recurrent problem statements, data acquisition, data management, analyse and report phases. The strength of qualitative research is its ability to provide complex textual descriptions of how people experience a given research issue. It includes information about a problem's "human-side ' with often contradictory behaviours, beliefs, opinions, emotions, and relationships.

Findings from qualitative data can often be extended to people with similar characteristics to those in the study population. However, gaining a rich and complex understanding of a specific social context or phenomenon takes precedence over eliciting data that can be generalised to other geographical areas or populations. In this sense, qualitative research differs slightly from scientific research in general.

Quantitative and qualitative research methods differ primarily in their analytical objectives, the types of questions they pose, the types of data collection instruments they use, the forms of data they produce and the degree of flexibility built into the study design. However, both research methodologies can be mapped (adapted from [3]) to a five-phase research workflow model: Problem Statement, Data Acquisition, Data Management, Analysis and Report.

Figure 1 presents the general research workflow model adapted from [3]. Blue and yellow arrows show the transitions among research phases. Blue arrows show the quantitative research workflow. Qualitative research, on the other hand, follows a process that can occur through the blue and yellow arrows interchangeably. Cycles and iterations are highlighted in each phase of the investigation workflow. In each phase, yellow emphasis is given to some relevant features of qualitative research.
The five phases of the research workflow are:
1. *Problem statement*: a research group composed of junior and senior scientists reviews theories and establishes or adjusts a set of research questions associated with a specific object of study; following these questions, the most appropriate research methodologies are defined, and hypotheses are identified, or in qualitative research, the theoretical framework of work is defined.
2. *Data acquisition*: is done using different instruments such as interviews, focus groups, questionnaires, and survey results. The collected data is explored and cleaned in successive iterations until it is guaranteed that the available data is reliable for the researchers' criteria.
3. *Data management*: in this recursive phase the acquired data is prepared to be used in experiments or reflections. An appropriate way of representing and manipulating the raw data is defined, the metadata obtained so far is evaluated and improved to reflect the context of use of the data. New metadata is produced for different research products (for example, interview transcripts, researchers' logs, and derived data).
4. *Analysis*: several rounds of experiments and measures in pure quantitative research are performed. Rounds of debate and reflections among researchers in pure qualitative research are instrumented. In this context, Hybrid research produces mixed research data applying both methods.
5. *Report*: this phase includes exploring, integrating and aggregating the observations and results into reports that can include textual, plots and visualisation metaphors adapted to multimedia documents. The voice and point of view of the researchers in the editors' role are relevant in qualitative research.

The role of humans across the different phases of qualitative research is mandatory, including in collecting data. So, part of a data-driven qualitative or a hybrid project is to choose the research team and their tasks and define strategies to screen informants, usually members of specific populations.

B. Qualitative Data Analysis Software

Several software tools that apply statistical techniques and machine learning algorithms are available for qualitative researchers. Woods et al. [4] state that Computer-Assisted Qualitative Data Analysis Software (CAQDAS) is a well-known tool for qualitative research. These tools support qualitative techniques and methods for applying Qualitative Data Analysis (QDA). ATLAS.ti [5], Dedoose [6], MAXQDA [7], NVivo [8] implement the REFI-QDA standard, an interoperability exchange format.

a) Computer-Assisted Qualitative Data Analysis Software (CAQDAS): Coding explores qualitative data in a systematic order by grouping, segregating, and thematically sorting to construct meaning [9]. CAQDAS allows researchers and practitioners to perform annotation, labelling, querying, audio and video transcription, pattern discovery, and report generation. Furthermore, CAQDAS tools provide different functionalities, such as the creation of field notes, thematic coding, search for connections, creation of memos (thoughtful comments), contextual analysis, frequency analysis, word location and data analysis presentation in different reporting formats [10].

b) Standard Data Exchange Format.: The REFI-QDA (Rotterdam Exchange Format Initiative) [1] standard allows the exchange of qualitative data to enable reuse in QDAS [11]. QDA software such as ATLAS.ti [5], Dedoose [6], MAXQDA [7], NVivo [8], QDAMiner [12], Quirkos [13] and Transana [14] adopt REFI-QDA standard.

C. Data curation

Data sets at an early collection stage are generally not ready for analysis or preservation; thus, extensive pre-processing, cleaning, transformation, documentation, and preservation actions are required to support usability, sharing, and preservation over time [15]. Data curation describes the actions required to maintain and use digital raw data throughout its life cycle for current and future interested users [16], [17]. Palmer et al. [18] define data curation in their article Zorich's Museum Management and Curatorship [19]. Data curation is formulated in the context of scientific data in [20]. Several data curation definitions are proposed in the literature [21]–[24], point out data curation as a process along the project lifecycle. Based on Choudhury and Huang's work [3], our view of research data curation considers the research workflow process to capture research practices and provenance of the dynamic nature of research data through an enriched research workflow model. Different types of metadata are associated with each type of research data produced in each research project phase. Operational metadata is related to the transition process among research phases (see Figure 1).

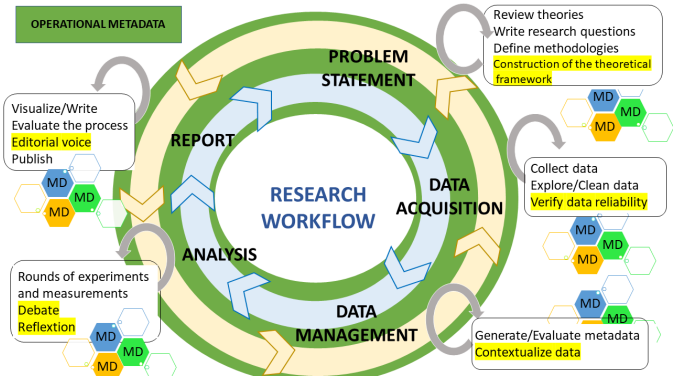

Fig. 1. Enriched research workflow model.

Data curation cannot be fully automated as it is often ad-hoc to a given type of context [22]. Moreover, semi-automatic curation processes can be a bottleneck for accessing data for developing a specific study [25]. According to Freitas and Curry [26] data curation processes can be categorised into different activities such as content creation, selection, classification, transformation, validation, and preservation. Hybrid human-algorithmic and HITL data curation approach, according to Freitas [26], are critical components for improving the automation of complex curation tasks. Several tools and systems are focus on the human-in-the-loop aspect of data science [27]. Human-in-the-loop systems allow human curators to avoid or reduce the amount of time spent on repetitive and automatable tasks, while keeping the entire curatorial process under human supervision [28].

D. Rationale

Hybrid methodologies combining quantitative and qualitative approaches produce heterogeneous research data that contain (i) the criteria, rules, hypothesis and research questions guiding the project and the participating scientists, (ii) heterogeneous multimedia documents collected using ad-hoc tools associated with observations and findings produced through the analysis phases; (iii) quantitative views of the content of research data produced by applying statistics, natural language analysis techniques, and numerical and artificial intelligence methods. The main characteristic of this type of project is that produced research data and knowledge are tagged with metadata, including provenance that keeps track of the conditions in which they are made.

Curation approaches [29], [30] and tagging tools provide partial automatic solutions for managing research data and associated knowledge. Existing curation solutions have mainly proposed approaches to extracting structural and quantitative metadata, vocabularies for textual documents, or elements of interest when they adopt linguistic perspectives used in libraries. None of these tools keeps track of research data versions to show the evolution of the changes focusing on the reasons/debates that produce them, thereby curating the

[1]https://www.qdasoftware.org

process of proposing findings and knowledge. Finally, the need for tools that allow semi-automating the research process with humans in the centre is critical for hybrid research projects.

Human intervention should guide automatic solutions generated by algorithms for curating and exploring data collections, inducing observations and interpretation of specific phenomena. Our approach considers the characteristics and requirements of research projects applying hybrid methods. We propose a human-in-the-loop approach for curating the research data. First describing the project working framework, including the role of its members in deciding the design of the project. Next, describing the evolution of research data and findings produced along the phases project and the metadata describing the decisions that trigger such evolution.

III. Curation environment for hybrid research projects data with HITL

The principle of the environment is to generate meta-data produced under HITL processes and a versioning approach to curate research data produced along a hybrid research project. First, meta-data includes technical aspects like the size of the files, format, provenance, production date and version. Next, structural meta-data like attributes and their types of tabular or semi-structured data, colours, distribution of images, number of scenes/second in a video and its duration. Finally, content meta-data including topic, descriptions, classification categories and objects in images. Providing, in this context, an interface to explore curated research data to allow technical decision-making about the project's progression and history, keeping track of the processes leading to specific decisions, conclusions and results.

Figure 2 shows the general architecture of the curation environment divided into three layers: storage, view management, and data processing. The *storage layer* provides persistence offered by services that archive raw data, data management systems that can store meta-data and versioning services well adapted for storing code implementing scripts that gather HITL interaction. Services in this layer interact to integrate data produced along a hybrid research project. The *data processing* layer consists of services for (i) harvesting data that are specialised according to the type of harvesting tool adopted (interview, questionnaire, etc.); (ii) data processing services that can extract meta-data according to the type of harvested data, and (iii) data analysis services that implement machine learning models, statistics, numerical models and ad-hoc processing operators. The *view management layer* implements the proposed curation model. The model is based on the notion of view for describing, organising and managing the data produced along a hybrid research project. The views have associated versioning operators that let scientists manipulate (create, update, version) data and explore them for making reports integrating findings in the different stages of the project.

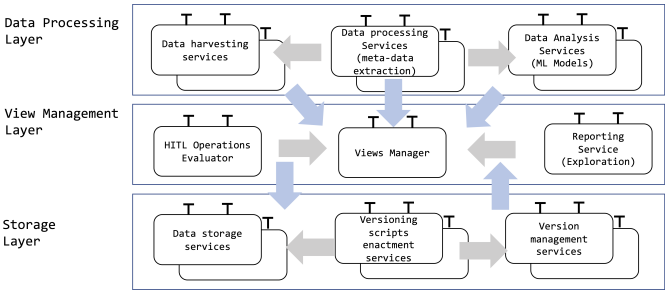

Fig. 2. Curation environment general architecture.

A. HITL Hybrid Research Curation Workflow

The main characteristic of hybrid research methods consists of non-linear incremental iterations. Therefore, we propose a data curation process of five recurrent stages without a predetermined beginning order phase. Phases are performed once or several times depending on the requirements and decisions made by scientists according to the study's intermediate results and problem statements. The phases of a hybrid research curation workflow follow the stages of the research workflow as a continuum (adapted from [3]). In each research project phase, researchers can use artificial intelligence tools to produce quantitative data or perform quantitative analysis. In our proposal, curating the Human-in-the-loop interventions means registering, maintaining, and contextualising decision-enhancing interactions between researchers and artificial intelligence tools.

B. View model of a hybrid research project

The notion of view is the central concept of our data model adapted for curating the content produced during a hybrid research project. The content generated within such type of project can have qualitative, quantitative, organisational, and document views (see Figure 3). Each view provides concepts for modelling the meta-data used for describing the content from different perspectives. Any entity in the views of the model has two statuses, undergoing and validated. Once validated, an entity cannot be modified; it can only have a new associated version.

a) Qualitative view: the central concept of this view is the concept of Tag which represents a multimedia document associated to any document produced within the research project. A tag is authored by a member of the project, it is time-stamped and it can have also associated tags produced by the same or different authors.

b) Quantitative view: gathers the statistical representation of the content of all documents of the other views. In can include frequency matrices and inverted indexes for representing the content of textual documents, statistics about the topics addressed in the documents, the type of comments produced by the members of the organisation.

c) Organisational view: represents the organisation of the human resources participating in the project, with their role and hierarchy if any or their type of expertise.

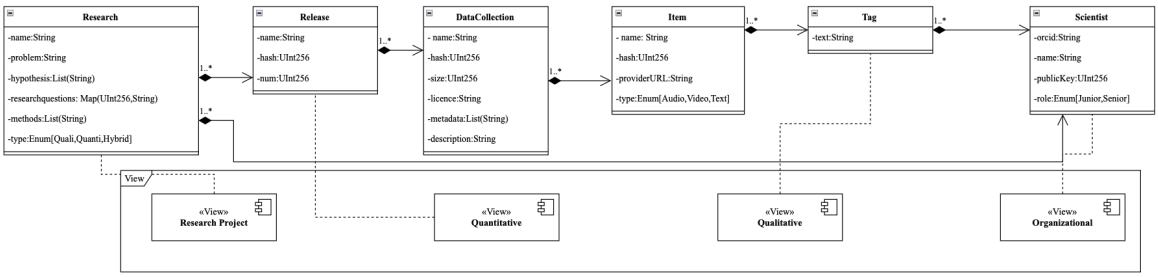

Fig. 3. UML class diagram of the view model of a hybrid research project.

The main concepts of the model is Scientist that can be junior or senior. Any member of the project organisation is entitled to produce documents and validate them. Thus any member has associated documents that she/he has produced, modified, and validated.

d) Research project documents: represent documents that serve as a basis for developing a hybrid research project. Documents are temporally, and they can be geographically tagged. They also have other associated provenance metadata, including the individuals that propose, produce, tag and modify them. *Research questions and hypothesis* are documents provided by individuals (scientists) willing to study a specific object. One or several questions can be associated with data harvesting document templates that will be used to collect data that can be used for answering questions.

One or several hypotheses can be associated with one or several research questions. *Data harvesting document* template describes tools like questionnaires, forms, and interviews used for harvesting data. *Raw data harvesting* result releases are multimedia responses given to target types of harvesting documents, filled in online or by a member of a qualitative research project on a given date. The document contains the answers given explicitly by a human or group of humans (sources), or it can be collected implicitly without the knowledge of the observed provider (for example, cameras observing people evolving in urban spaces). In each of the framework's stages and phases, collecting and exploring raw data begins a curation process with HITL. We propose that the researcher intervenes in the loop in each iteration where ML algorithms are applied. In this intervention, the researcher commits the rationale of the changes in the respective branch of the version-controlled system. The objective of these approach is to go beyond data curation, and rather curate only the research project, establish the traceability decision making process within the project.

C. Managing qualitative and quantitative views

To manage qualitative and quantitative views, we propose a versioning approach, documents produced during a hybrid research project to keep track of incrementally-different versions of information (collected data, harvesting tools, research questions, observations and tags). Each stage of a research process can be identified in independent branches and modelled with a partially ordered set. For example, research questions can be registered in a branch x, hypotheses in a branch y and the methods in a branch called z. Finally, merging each branch towards a final version called $\{x, y, z\}$ keeps track of all research artefacts modifications along the research project. Currently, versioning documents, ML algorithms and data science pipeline stages can also be registered in the repository log. In these research branch versioning approach documents produced during a hybrid research project keep track of incrementally-different versions of information (collected data, harvesting tools, research questions, observations and tags). The documents are time-stamped and have a unique ID and a status (undergoing, validated). The time-stamp and the ID and status compose an internal version identifier that may evolve many times. Then, validated documents have a release version that changes far less often. Inspired by version managing operators, we propose a set of operators devoted to managing documents of the different phases of the project. Some of them can only be used by a user with specific roles in the organisation of the project.

The following operations aims to manage the general hybrid research project:

- `init` creates a new, empty curation repository for a research project.
- `clone url` retrieve an entire existing project with existing branches and commits.
- `status` shows the actual status of the research repository.
- `branch name` creates a branch named name.
- `merge branchname` a single branch's into the current research branch.
- `checkout branchname` will change to another branch's into the current research branch.

The following operators allow members to produce new documents during a phase of the research workflow to manage them, organise them in files, and create versions (undergoing and committed versions that the overall research team can validate).

- `add` adds a new file to the repository, prior to performing the corresponding commit.
- `add remote url` link a new file from an external resource URL.

- `mv` moves a file from one directory to another (or renames it).
- `commit -m "research message"` changes the status of the entities created by a user to "validate" in the current branch and start version the artifacts.
- `pull url` pull from the tracking remote branch url and updates a working copy with changes from the repository possibly performed by other members of the project.
- `push url` push actual branch to the remote branch url.

The operators are selected to simplify operations currently used by versioning tools like CVS, SVN and GIT. In particular, the objective is to provide agnostic operations of mentioned technologies. To make the project progress, scientists must analyse comments and observations, generating reports that can include quantitative content. The following fine-grained operators can be combined into scripts, notebooks or domain-specific SQL or GraphQL [31] like exploration languages to produce reports about the results and milestones:

- `ls` allows to see a list of files in a repository without creating a working copy.
- `diff` reveals the differences between a working copy and the copy in the master.
- `status` prints the status of working copy files and directories.
- `info` displays information about an configuration element.
- `tag id` tags the current version.
- `log` shows the series of actions performed on top of a document and associated provenance information.

In this context, one of the main contributions of this work is a curated view model within a version control system strategy for managing qualitative and quantitative research data. In this context, a team of researchers can identify in their corresponding branches each of the changes in their research questions, hypotheses, methods and rationale of data analysis.

IV. Use Case: Curating the Analysis of Graffiti Narratives in the Old Downtown of Montevideo

Consider the use case of a hybrid research project willing to analyse graffiti narratives in the old downtown of Montevideo [1]. Figure 6 shows how our view model can be instantiated throw three iteration stages. In this use case example, qualitative researchers highlight research context associated with cultural aspects, specific communication of symbols and appropriation of spaces in photo murals throughout the city. In these murals, the typography, signatures, styles, and strokes are a substantial part of the study that will be "tagged" and codified later. This coding allows researchers to establish a classification of the elements to create, through iterations, a grounded specific theory from research field findings.

a) First stage. Research project preparation: The project begins with iterations on the problem statement phase. Reviews on literature, identification of a set of research questions, and the definition of inclusion/exclusion criteria of the data collection method. In this use case, the answer to a social, cultural or ethnographic aspect begins with a research question, such as: "What types of sociocultural communication are present in the photo murals of Montevideo?" (see Fig. 6 *GraffitiProject::Research* instance). Sociologist and anthropologists (seniors and juniors) are part of the research team, with their roles associated and experience. The organisation of the team is gathered by instances of the classes of the organisational view of our model. Hybrid research method selection, hypothesis and geographical field location of old Montevideo downtown are defined early. Codebook notes guidelines are part of the expert decision that senior researchers propose. The students' registration of the photos is made, with a stated junior role defined in the organisational view. Inclusion and exclusion criteria are determined by all the research team (juniors and seniors) as stated in the roles defined by the organisational view of our model. For example, graffiti that is part of the shop is excluded from those that refer to cultural or political aspects.

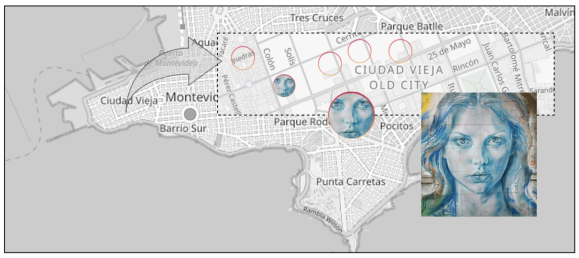

Fig. 4. Harvested graffiti in Montevideo's old downtown during phase 1 (adapted from [1]).

In a second iteration, researchers focus their study on a specific area, where the political aspects and expressions of protest have a more significant influence in the old downtown city Montevideo (proximity of the government house). Open Streets Maps (OSM) API allows researchers to georeference graffiti. Researchers start with a subset of categories. These categories define the initial selection criteria for harvesting photos. Examples of narrative tags in coding images are: marginality, anonymity, spontaneity, scenicity, speed of stroke execution, precariousness, and transience (see Fig. 6 *Tag1::Tag* instance). They also use predefined categories for classifying graffiti: Tags, Throw-up, Pieces, Widestyle, 3D, top to Bottoms, whole cars, and whole trains fading.

b) Second stage. Data Acquisition: Once the research group begins the work in the field, georeferenced and spatio-temporal photos are taken in the City. Figure 4) shows an overview of the graffiti taken by [1] in Montevideo's old downtown chosen in phase 1. Figure 5) shows Harvested graffiti in Montevideo's old downtown tagged in social media networks [1]. These raw material collections are the primary resource for data harvesting procedures. The team members tagged the harvested pictures spontaneously with qualitative and quantitative meta-data. For example, tags like "urban inscription", "smoke", and "reflection" are classified manually according to the following categories: anonymity, spontaneity

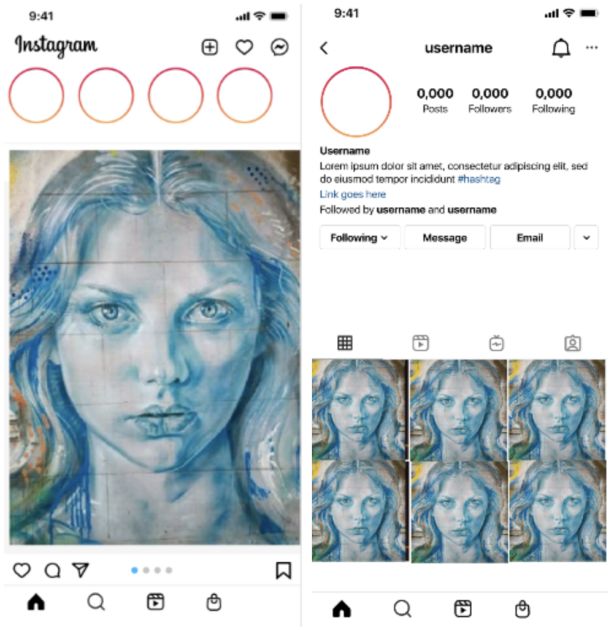

Fig. 5. Harvested graffiti in Montevideo's old downtown tagged in social networks (adapted from [1]).

and staging. Tags are gathered as instances of the class *Tags* of the model's view *Qualitative*. The images are also processed automatically to classify them using ML algorithms such as "pop art" or "cubism". These classification results are also gathered as tags, instances of the class *Tag* of the model's view *Quantitative*. Researchers can analyse raw material and data harvesting preliminary results using the operators `ls` , `diff`, `log` and `info`. According to the results, the research team can adjust the initial research question, harvesting criteria, urban spaces to visit or even tagging criteria.

c) Third stage. Adjusting the research question and studying findings: This phase begins with a reflexivity component performed by the team. The research question, hypothesis and criteria adopted in the first phase may change. When researchers identify that it is more interesting to reformulate the question: "What types of political communication are present in the photo murals of Montevideo?". Research questions changes can be impacted in the version control system with `pull`, `add`, `commit`, and `pull` operators. Also, each researcher could previously create a new `branch` in case of working simultaneously to perform `merge` and `pull` operators later.

d) Fourth stage. Analysis: In this stage, georeferenced and tagged photos of the Old downtown City of Montevideo are automatically processed by ML algorithms. HITL feedback appears to register the rationale of the research approach. Feedback from ML algorithms in the HITL loop is reflected in the processing feedback module and the researchers' branch. The criteria used to make decisions and adjustments are also gathered in documents. New harvested pictures, observations and interpretations with `diff` and `log` operators can then be compared with previous versions and findings. This iteration's set converges on a consensus of the criteria to follow in the next stages throughout a spiralling process of analysis and reflection.

e) Fifth stage: Reporting: Reports generated within qualitative, quantitative, organisational and document views are available to researchers. In the qualitative view, the research team can group and visualise tagging, keywords and meta-data of the photos and comments —for example, political classifications, contra-cultural expressions or political propaganda. The quantitative view summarises the numerical data —for example, georeferenced spaces, 3D style graffiti, statistics and natural language analysis results. In the organisational view, researchers can identify the role and fundamentally what decision-making rationale was made throughout the research process by creating artefacts and validating them. Finally, the research project document view reports all questionnaires, forms, interview scripts, and relevant documents of the research process.

A. Lessons learned

RQ_1: What is research data curation from a hybrid (qualitative and quantitative) perspective?
We define hybrid research projects' data curation as "the process of identifying, systematizing, managing and versioning research data produced along the project stages". Hybrid research methodologies call for data curation strategies that keep track of the data that describe the tools, strategies, hypothesis, and data harvesting criteria, defined a priori by a scientific team. They call for curating the data produced during the target study, the development process, and project milestones that are non-linear and depend on decision-making stages. They also call for quantitative methods to complete the findings and confront them with complementary visions produced automatically. The use-case about graffiti shows the complexity of the curation process with many different details to be considered to keep valuable and critical track of the whole research process. In this context, we propose a versioning system that allows recording decision-making throughout the research process.
RQ_2: What software tools are adapted for hybrid (qualitative and quantitative) research projects? Curation environments adapted for hybrid research must be flexible and provide toolkits adapted for managing (modelling, storing and exploring) heterogeneous data regarding the project's design and data produced by humans and automatically during the life-cycle stages. The adapted tools depend on the type of project, so flexible architectures like service-based ones allow to extend the tools and the data types. The environments must be interactive and allow scientists to intervene at any curation stage and operation, including automatic processes.

V. Conclusions and Future Work

This paper proposes a curation environment ad-hoc for managing content produced during hybrid research projects. The environment is based on a view model that we propose for representing meta-data produced across the phases of a hybrid

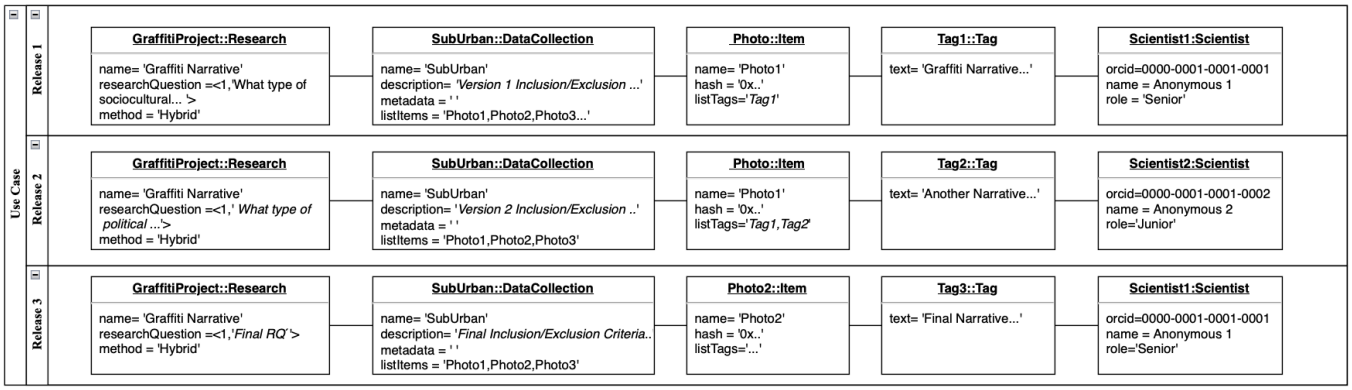

Fig. 6. Use case example.

research project. The meta-data is produced under HITL processes and a versioning approach. Therefore, we redefine versioning operators to manage hybrid research documents' progression.

Our current work includes validating our approach in the context of the project MENTOR (seMantic Exploration aNd curaTion of Open hybrid Research), supported by the National Agency for Research and Innovation (ANII), in Uruguay, which proposes a set of use cases for dealing with three hybrid research studies in anthropology, education and ethnography.

REFERENCES

[1] C. G. Romeo, "El graffiti minado, análisis cuantitativo aplicado," Master's thesis, Facultad de información y comunicación, Universidad de la República (UdelaR), 2021, to appear.

[2] E. Dragut, Y. Li, L. Popa, and S. Vucetic, "Data science with human in the loop," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 4123–4124.

[3] G. S. Choudhury, C. Huang, and C. L. Palmer, "Updating the dcc curation lifecycle model," *Int. J. Digit. Curation*, vol. 15, pp. 1–12, 2020.

[4] M. Woods, R. Macklin, and G. K. Lewis, "Researcher reflexivity: exploring the impacts of caqdas use," *International Journal of Social Research Methodology*, vol. 19, no. 4, pp. 385–403, 2016.

[5] ATLAS.ti, "ATLAS.ti," https://atlasti.com, last accessed July 2022.

[6] Dedoose, "Dedoose," https://www.dedoose.com/, last accessed July 2022.

[7] V. Software, "Maxqda," http://maxqda.com, last accessed July 2022.

[8] NVivo, "Nvivo," https://www.qsrinternational.com/, last accessed July 2022.

[9] N.-C. Chen, M. Drouhard, R. Kocielnik, J. Suh, and C. R. Aragon, "Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 2, pp. 1–20, 2018.

[10] J. C. Evers, "Current issues in qualitative data analysis software (qdas): A user and developer perspective," *The Qualitative Report*, vol. 23, no. 13, pp. 61–73, 2018.

[11] S. Karcher, D. D. Kirilova, C. Pagé, and N. Weber, "How data curation enables epistemically responsible reuse of qualitative data," *The Qualitative Report*, vol. 26, no. 6, pp. 1996–2010, 2021.

[12] QDAMiner, "Qdaminer," https://provalisresearch.com/products/qualitative-data-analysis-software/, last accessed July 2022.

[13] Quirkos, "Quirkos," https://www.quirkos.com, last accessed July 2022.

[14] Transana, "Transana," https://www.transana.com, last accessed July 2022.

[15] S. Lafia, A. Thomer, D. Bleckley, D. Akmon, and L. Hemphill, "Leveraging machine learning to detect data curation activities," in *2021 IEEE 17th International Conference on eScience (eScience)*. IEEE, 2021, pp. 149–158.

[16] I. G. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino, "Data wrangling: The challenging yourney from the wild to the lake." in *CIDR*, 2015.

[17] C. Huang, J.-S. Lee, and C. L. Palmer, "Dcc curation lifecycle model 2.0: Literature review and comparative analysis," *'Int.Journal of Digital Curation'*, vol. 15, no. 1, 2020.

[18] C. Palmer, N. M. Weber, A. H. Renear, and T. Muñoz, "Foundations of data curation: The pedagogy and practice of purposeful work with research data," *Archives Journal*, vol. 3, 2013.

[19] D. M. Zorich, "Data management: Managing electronic information: Data curation in museums," *Museum Management and Curatorship*, vol. 14, no. 4, pp. 430–432, 1995.

[20] A. Lyngdoh, "What we leave behind: the future of data curation," in *Trends, Discovery, and People in the Digital Age*. Elsevier, 2013, pp. 153–165.

[21] F. Zouari, C. Ghedira-Guegan, N. Kabachi, and K. Boukadi, "Towards an adaptive curation services composition based on machine learning," in *2021 IEEE International Conference on Web Services (ICWS)*. IEEE, 2021, pp. 73–78.

[22] S. Thirumuruganathan, N. Tang, M. Ouzzani, and A. Doan, "Data curation with deep learning [vision]," *arXiv preprint arXiv:1803.01384*, 2018.

[23] F. A. Sposito, "What do data curators care about? data quality, user trust, and the data reuse plan," *'http://library.ifla.org/id/eprint/1797/1/S06-2017-sposito-en.pdf'*, last accessed July 2022.

[24] P. C. Arocena, B. Glavic, G. Mecca, R. J. Miller, P. Papotti, and D. Santoro, "Benchmarking data curation systems." *IEEE Data Eng. Bull.*, vol. 39, no. 2, pp. 47–62, 2016.

[25] D. Garat and D. Wonsever, "Automatic curation of court documents: Anonymizing personal data," *Information*, vol. 13, no. 1, p. 27, 2022.

[26] A. Freitas and E. Curry, "Big data curation," in *New horizons for a data-driven economy*. Springer, Cham, 2016, pp. 87–118.

[27] Z. Shang, E. Zgraggen, B. Buratti, F. Kossmann, P. Eichmann, Y. Chung, C. Binnig, E. Upfal, and T. Kraska, "Democratizing data science through interactive curation of ml pipelines," in *Proceedings of the 2019 international conference on management of data*, 2019, pp. 1171–1188.

[28] M. J. Weber N., Karcher S., "Open source tools for scaling data curation at QDR," *The code4lib journal*, 2020.

[29] G. Vargas-Solar, G. Kemp, I. Hernández-Gallegos, J. Espinosa-Oviedo, C. da Silva, and P. Ghodous, "Exploring and curating data collections with curare," in *Proceeding of the 35ème Conférence sur la Gestion de Données–Principes, Technologies et Applications*, 2019.

[30] G. Vargas-Solar, J.-L. Zechinelli-Martini, and J. A. Espinosa-Oviedo, "Enacting data science pipelines for exploring graphs: from libraries to studios," in *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*. Springer, 2020, pp. 271–280.

[31] A. Jobst, D. Atzberger, T. Cech, W. Scheibel, M. Trapp, and J. Döllner, "Efficient github crawling using the graphql api," in *International Conference on Computational Science and Its Applications*. Springer, 2022, pp. 662–677.