MuseIT
MuseIT
Multisensory, User-centred, Shared cultural
Experiences through Interactive Technologies
2022 - 2025

IPR Management through data trust and provenance technologies

Document Information

Project Name:	Multisensory, User-centred, Shared cultural Experiences through Interactive Technologies
Project Acronym:	MuseIT
Grant Agreement No:	101061441
Deliverable No:	D6.2
Deliverable Name:	IPR Management through data trust and provenance technologies
Work Package:	WP6
Task:	T6.3
Dissemination Level:	PU
Deliverable Type:	R
Lead Organization:	KCL
Lead Member:	Albert Meroño Penuela
Submission month:	M30
Date:	31 March 2025

Co-funded by
the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Document history

Revision	Date	Description / Reason of change	Submission by
v0.1	31-01-2025	Structure proposal and initial draft	Nitisha Jain
v0.2	14-02-2025	First draft for internal review	Nitisha Jain, Albert Meroño
v0.3	03-03-2025	Second draft addressing review comments	Nitisha Jain
v0.4	20-03-2025	Final draft addressing the PMB review comments	Nitisha Jain
v1.0	27-03-2025	Final draft submitted to the EU	Renata Sadula

Authors

Partner	Name(s)
KCL	Albert Meroño Peñuela
KCL	Nitisha Jain
DANS	Vyacheslav Tykhonov

Contributors

Partner	Contribution type	Name
EXUS	Review	Roberto Maffulli
DANS	Review	Andrea Scharnhorst
CTL	PMB Review	Stelios Kontogiannis
KCL	Contributor	See above

Glossary

Acronym	Definition
AFC	Automated Fact-Checking
DCAT	Data Catalog Vocabulary
DRM	Digital Rights Management
DUO	Data Use Ontology
FAIR	Findability, Accessibility, Interoperability, and Reusability
IPR	Intellectual Property Rights
LOD	Linked Open Data
MFC	Multimodal Fact Checking
ML	Machine Learning
OCR	Optical Character Recognition
RAI	Responsible AI

Table of contents

Executive Summary

This deliverable presents an integrated framework combining metadata standards and multimodal verification techniques for managing IPR and provenance of MuseIT's digital cultural heritage assets. The goal is to create a structured and automated system to protect, authenticate, and ensure the ethical use of digital artifacts. The focus is on managing the IPR of digital modalities and their provenance, addressing the specific complexities involved in handling digital cultural heritage assets, such as maintaining authenticity, ensuring proper attribution, and preventing unauthorized reproduction.

The novelty of this approach lies in the combination of the Croissant[1] metadata standard and multimodal automated fact-checking (AFC). Croissant enhances dataset discoverability and interoperability, enabling seamless integration with machine learning workflows while ensuring traceability and legal compliance. Multimodal AFC validates the integrity of images, videos, and textual descriptions through claim detection, evidence retrieval, and manipulation detection—ensuring that digital assets are accurately represented and protected from misinformation and unauthorized use.

The intersection of structured metadata (Croissant) and multimodal fact-checking offers a novel and holistic approach to IPR management. By combining these two technologies, MuseIT bridges the gap between fundamental research on dataset documentation and real-world application needs in the digital cultural heritage sector. The requirements and achievements of MuseIT in multimodal representation so far have directly influenced this work demonstrating how the project's application area fosters research on fundamental challenges in digital content verification.

This deliverable is closely aligned with MuseIT's deliverables D1.2 Data Management Plan (DMP) and D8.1 Initial Exploitation Plan. D1.2 provided a foundational structure for handling data assets, ensuring compliance with FAIR principles and best practices in metadata standards. The methodologies in D6.2 build upon these principles by integrating provenance tracking and trust mechanisms into the metadata framework. Additionally, the Initial Exploitation Plan (D8.1) outlines strategies for commercializing and sustaining MuseIT's outputs. The alignment between D6.2 and D8.1 ensures that provenance and IPR management strategies contribute to the broader objectives of asset exploitation and long-term sustainability.

Overall, by integrating Croissant and multimodal fact checking, MuseIT delivers a robust, transparent, and scalable solution for managing IPR in digital cultural heritage, ensuring that cultural assets remain accessible, trustworthy, and protected in an increasingly digital world.

[1] https://mlcommons.org/working-groups/data/croissant/

1. Introduction

The increasing digitization of cultural heritage assets presents significant challenges in IPR management, authenticity verification, and digital rights enforcement. Traditional text-based metadata solutions are insufficient for handling the complexity of multimodal digital content, which includes images, videos, and textual descriptions. Misrepresentation, manipulation, and unauthorized use of these assets threaten their integrity and legal protection. For example, altered historical photographs or misattributed artworks can distort historical records and mislead the public.

This deliverable brings together two major contributions: the Croissant metadata standard and multimodal fact-checking. These approaches provide complementary solutions for ensuring the provenance, authenticity, and responsible use of digital cultural heritage assets.

To address these issues, structured approaches leveraging metadata standards and automated verification mechanisms are essential. In our recent work, we introduced "Croissant: A Metadata Format for ML-Ready Datasets," published in the Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2024) [1]. This paper presents Croissant, a standardized metadata format designed to enhance dataset discoverability, interoperability, and responsible AI integration. By providing structured descriptions for multimodal assets, Croissant supports provenance tracking, licensing transparency, and compliance with IPR regulations.

Additionally, we explored multimodal fact-checking frameworks in our survey titled "Multimodal Automated Fact-Checking: A Survey," published in the Findings of the Association for Computational Linguistics, EMNLP 2023 [2]. This work conceptualizes a framework for automated fact-checking that includes subtasks unique to multimodal misinformation. **We focus on four modalities prevalent in real-world fact-checking: text, image, audio, and video. By combining Croissant's structured metadata with fact-checking techniques, cultural heritage assets can be effectively tracked, authenticated, and safeguarded against unauthorized alterations.** This point is central to our approach: structured metadata alone is not enough—multimodal fact-checking is necessary to ensure the integrity of digital assets.

The MuseIT project aims to co-design, develop, and co-evaluate a multisensory, user-centered platform for enriched engagement with cultural assets, with inclusion and equal opportunity for all as core principles. By integrating Croissant's standardized metadata format, MuseIT can enhance the discoverability and interoperability of its digital cultural heritage assets, ensuring that diverse user groups, including those with disabilities, can access and engage with the content effectively.

Furthermore, the multimodal fact-checking framework provides a robust mechanism for verifying the authenticity of digital assets. By applying this framework, MuseIT can ensure that the cultural heritage assets it develops and shares are accurate, trustworthy, and protected against unauthorized alterations, thereby maintaining the integrity and legal protection of these assets.

This report builds upon our previous works, aiming to provide a comprehensive solution for managing, authenticating, and protecting digital cultural heritage assets. Previous EU projects, such as the Polifonia[2], has focused on extracting and structuring license information from web resources, using semantic web technologies to represent licensing terms as ontologies and knowledge graphs[3, 4]. This approach enabled querying and reasoning over license metadata but remained largely observational—

[2] https://polifonia-project.eu/

analyzing existing datasets rather than modifying them. In contrast, our work takes a more interventional role by proposing mechanisms to enrich ML dataset metadata (via Croissant) and support multimodal fact-checking (MFC). Rather than solely documenting licensing practices, we provide tools to enhance dataset interoperability and ensure responsible reuse, making our approach complementary to, but distinct from, Polifonia's. By integrating Croissant's metadata framework with multimodal fact-checking methodologies, we propose a unified approach to address the challenges of IPR management and authenticity verification in the digital age.

The rest of this report is structured as follows: Chapter 2 introduces the Croissant metadata format, detailing its structure and evaluation. Chapter 3 discusses multimodal fact-checking as a framework for provenance verification and IPR protection. Finally, Chapter 4 presents the overall conclusions and future directions.

2.Croissant: A Metadata Format for ML-Ready Datasets

2.1 Introduction

Effective data management plays a crucial role in ensuring the quality and usability of datasets. Yet, working with data remains time-consuming and challenging due to a wide variety of data formats, the lack of interoperability between tools, and the difficulty of discovering and combining datasets [5, 6]. Metadata standards such as Data Catalog Vocabulary (DCAT), schema.org, and Data Packages provide essential metadata structures, yet they often lack specialized capabilities for machine learning integration and the complex needs of multimodal content. We have proposed Croissant, a metadata format designed to improve ML datasets' discoverability, portability, reproducibility, and interoperability.

Croissant addresses these challenges by combining comprehensive dataset documentation with ML-specific attributes that describe dataset splits, data augmentation techniques, and model-specific metadata. By supporting structured descriptions for multimodal assets like images, audio, and video, Croissant improves dataset discoverability and portability. In addition, its Resource Layer offers flexible mechanisms for representing datasets with complex file structures, ensuring that multimodal content—including 3D models, audiovisual data, and text—is easily described and accessed.

Croissant also extends beyond standard metadata frameworks by introducing the Croissant-RAI extension [7], which captures responsible AI documentation features such as dataset biases, labeling processes, and provenance tracking [8]. This emphasis on responsible data management aligns with MuseIT's focus on ensuring digital cultural heritage data is ethically managed and accessible.

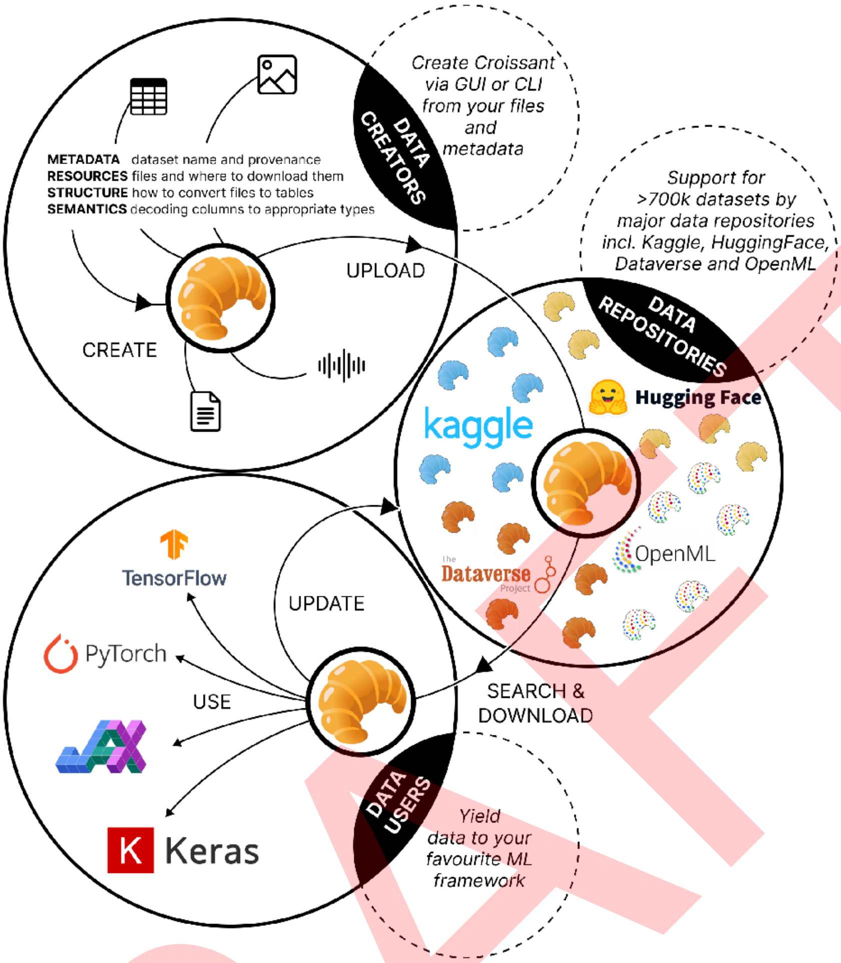
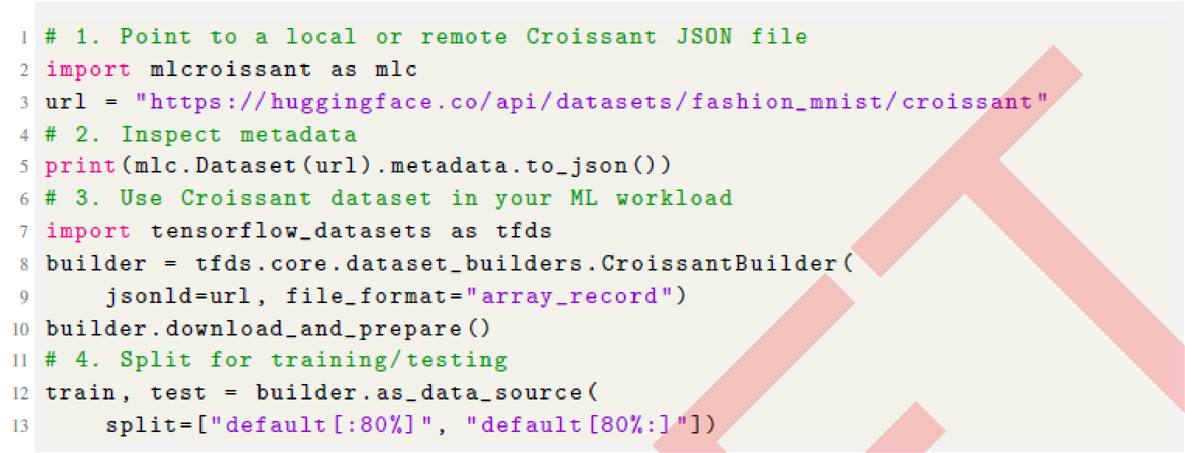


Figure 1: The Croissant lifecycle and ecosystem

In MuseIT, Croissant plays a pivotal role by ensuring that cultural heritage datasets are documented in a structured yet flexible format that supports diverse media types. This alignment empowers MuseIT to create ML-ready datasets that facilitate automated content generation, provenance tracking, and improved accessibility for users with disabilities. By adopting Croissant, MuseIT ensures that its cultural heritage data assets are not only well-documented but also interoperable with contemporary ML tools and repositories.

Figure 1 gives an overview of the Croissant lifecycle and ecosystem. Croissant makes datasets "ML-ready" by recording ML-specific metadata that enables them to be loaded directly into ML frameworks and tools (see Figure 2 for sample code). Croissant describes datasets' attributes, the resources they contain, and their structure and semantics. This uniform description streamlines their usage and sharing within the ML community and between ML platforms and tools while fostering responsible ML practices. Croissant can describe most types of data commonly used in ML workflows, such as images, text, audio, or tabular. While datasets come in a variety of data formats and layouts, Croissant exposes a unified "view" over these resources. It lets users add semantic descriptions and ML-specific information. The Croissant vocabulary [9] does not require changing the underlying data representation and can thus be easily added to existing datasets and adopted by dataset repositories.

To assess Croissant's usability, we conducted a preliminary usability evaluation on metadata creation for language, vision, audio, and multimodal datasets. Several practitioners annotated ten widely used ML datasets. We analyzed the consistency of their responses and collected their feedback on Croissant.



```python
# 1. Point to a local or remote Croissant JSON file
import mlcroissant as mlc
url = "https://huggingface.co/api/datasets/fashion_mnist/croissant"
# 2. Inspect metadata
print(mlc.Dataset(url).metadata.to_json())
# 3. Use Croissant dataset in your ML workload
import tensorflow_datasets as tfds
builder = tfds.core.dataset_builders.CroissantBuilder(
    jsonld=url, file_format="array_record")
builder.download_and_prepare()
# 4. Split for training/testing
train, test = builder.as_data_source(
    split=["default[:80%]", "default[80%:]"])
```

Figure 2: Users can easily inspect datasets and use them in data loaders with Croissant.

2.2. Related Work

While there have been many prior efforts in standardizing dataset metadata, they typically lack ML-specific support, do not work with existing ML tools, or lag behind the demands of dynamically evolving requirements, such as responsible ML. We outline the state of the field below.

2.2.1 Vocabularies for Dataset Documentation

Dataset documentation is indispensable for effective data management and serves as a foundational element for training and evaluating ML models [10]. Metadata descriptions of datasets enhance their discoverability, interoperability, and usability, which is critical for advancing research and data-driven applications. Ontologies and vocabularies are semantic web tools used to standardize dataset documentation. While vocabularies comprise sets of terms and their meanings to describe data consistently, ontologies provide a structured framework to define and relate these concepts within a domain. Ontologies and vocabularies are evaluated for their coverage (i.e., do they represent all relevant concepts), accuracy (correctness of definitions and relationships), consistency (no logical contradictions), and usability (ease of use and integration). This is done through methods like competency questions, expert validation, and use-case testing [11].

2.2.2 Standards for Catalogs and Metadata

With the increase of data availability online, various efforts have focused on making data both discoverable and user-friendly by supplementing datasets with comprehensive metadata. Such metadata includes details about the data, such as authorship, format, and intended use, all structured consistently to support automated processing and retrieval. Key efforts towards documentation have led to the creation of standards like the DCAT [12] and the Dataset vocabulary in schema.org [13]. DCAT facilitates interoperability among web-based data catalogs, enabling users to aggregate, classify, and filter datasets efficiently. Schema.org [14] acts as a de facto standard for metadata, helping search engines discover and index published web content, including datasets, thus enhancing dataset accessibility and understandability. This versatility allows schema.org to describe a wide array of content types effectively. Other frameworks, such as Data Packages [15] and CSV on the Web [16] support methods for describing and exchanging tabular data. The Global Alliance for Genomics and Health's Data Use Ontology (DUO) [17] refines data usage terms with optional modifiers, improving clarity in genomic data sharing agreements. Efforts to integrate FAIR principles (Findability, Accessibility, Interoperability, and Reusability) [18] metadata vocabularies are also noteworthy. Despite their utility for specific domains and formats, these standards do not entirely meet the specialized needs of data management within the ML domain.

2.2.3 Operationalizing Responsible AI through Data Work

Data-centric ML [5, 16] is increasingly seen as critical to the development of trustworthy ML systems, considering RAI aspects such as fairness, accountability, transparency, data privacy and governance, safety, and robustness [20]. Seminal works, such as Datasheets for Datasets [10] and Data Statements [21], have emphasized the importance of dataset documentation to assess and increase the trustworthiness of ML systems. Several related documentation efforts such as Data Cards [22] and Data Nutrition Labels [23] have been inspired them. ML data repositories, such as Kaggle [24], OpenML [25] and Hugging Face [26], have initiated their own metadata documentation efforts. Hugging Face, for example, provides Dataset Cards [27] that include summaries, fields, splits, potential social impacts, and biases inherent in the datasets.

These approaches typically rely on data documentation written in natural language, without a standard machine-readable representation, which makes data documentation challenging for machines to read and process. Croissant fills this gap by providing a standardized framework for data documentation that ensures semantic consistency and machine readability, thereby facilitating seamless integration with existing tools and frameworks used by the ML community.

2.3. The Croissant Format

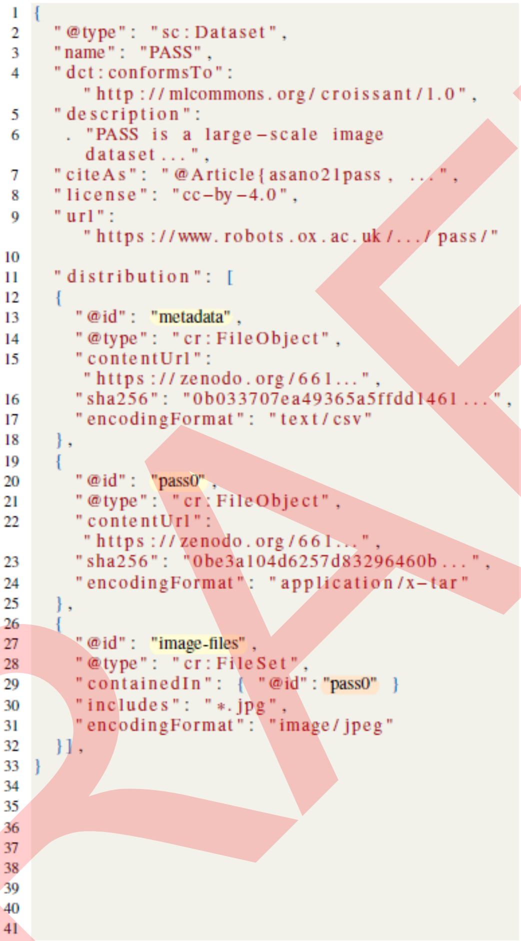
The Croissant format is a community-driven metadata vocabulary for describing datasets that builds on Schema.org. Croissant is divided into four layers: (i) The Dataset Metadata Layer, containing relevant information such as name, description, and version. (ii) The Resource Layer describes the source data used in the dataset. (iii) The Structure Layer, describing and organizing the structure of the resources. (iv) The Semantic Layer, which provides ML-specific data interpretation and semantics. A more detailed description of the Croissant format can be found in the official specification [28]. Documentation and code are available online[3].

In the remainder of this section, we illustrate each layer with examples from popular ML datasets. Afterwards, we briefly describe the Croissant Responsible AI extension, and then provide an overview of ML frameworks, tools, and repositories that currently support Croissant.

[3] https://docs.mlcommons.org/croissant/

2.3.1 The Dataset Metadata Layer

Croissant dataset descriptions, illustrated in Figure 3, are based on schema.org/Dataset, a widely adopted vocabulary for datasets on the Web [13], hence ensuring interoperability with existing standards and tools. Croissant specifies constraints on which schema.org properties are required, recommended, and optional, and adds additional properties, e.g., to represent snapshots, live datasets, and citation information.

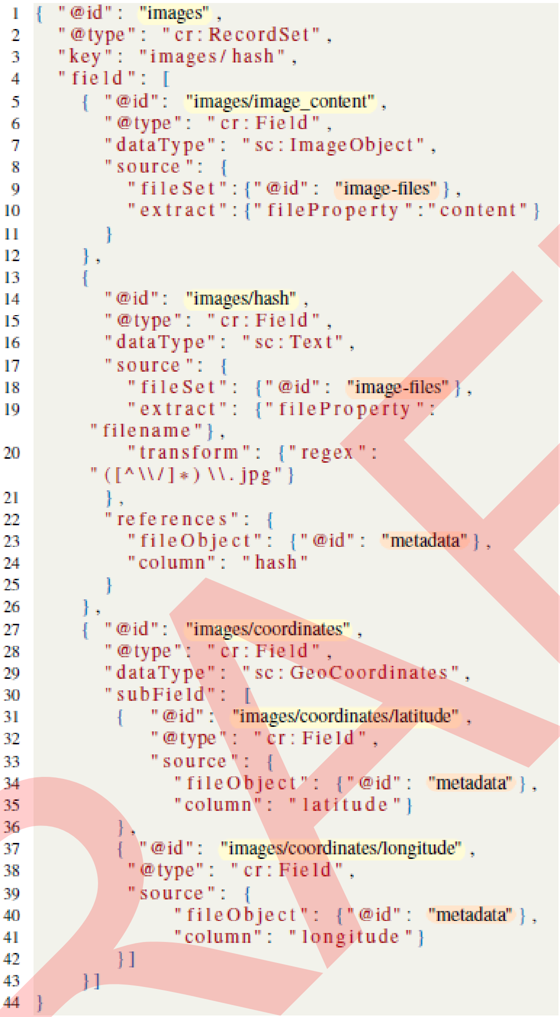


```
1  {
2    "@type": "sc:Dataset",
3    "name": "PASS",
4    "dct:conformsTo":
        "http://mlcommons.org/croissant/1.0",
5    "description":
6     . "PASS is a large-scale image
        dataset...",
7    "citeAs": "@Article{asano21pass, ...",
8    "license": "cc-by-4.0",
9    "url":
        "https://www.robots.ox.ac.uk/.../pass/"
10
11   "distribution": [
12   {
13     "@id": "metadata",
14     "@type": "cr:FileObject",
15     "contentUrl":
        "https://zenodo.org/661...",
16     "sha256": "0b033707ea49365a5ffdd1461...",
17     "encodingFormat": "text/csv"
18   },
19   {
20     "@id": "pass0",
21     "@type": "cr:FileObject",
22     "contentUrl":
        "https://zenodo.org/661...",
23     "sha256": "0be3a104d6257d83296460b...",
24     "encodingFormat": "application/x-tar"
25   },
26   {
27     "@id": "image-files",
28     "@type": "cr:FileSet",
29     "containedIn": { "@id": "pass0" }
30     "includes": "*.jpg",
31     "encodingFormat": "image/jpeg"
32   }],
33 }
34
35
36
37
38
39
40
41
```

Figure 3: Dataset metadata and resources for the PASS dataset.

2.3.2 The Resources Layer

This layer represents the data resources (e.g., files) of the dataset. Schema.org properties are insufficient to adequately describe dataset contents with complex layouts, which are common for ML datasets. This layer provides two primitive classes to address this limitation and describe dataset resources: FileObject to describe individual files and FileSet to describe sets of files. Figure 3 shows an excerpt of the Croissant definition of the PASS dataset [29], where declarations of object names are highlighted in yellow, with references in orange. This distribution includes two FileObjects: a CSV file containing metadata about the dataset (line 13) and an archive file containing images (line 20). Moreover, FileSet (in line 27) is used to refer to a collection of images, videos, or text files that contain the (unlabeled) data used for training and inference. Since there can be numerous files, FileSets are specified with inclusion/exclusion filters (e.g., a pattern matching all files that should be included) as shown on line 30.



```
1  { "@id": "images",
2    "@type": "cr:RecordSet",
3    "key": "images/hash",
4    "field": [
5      { "@id": "images/image_content",
6        "@type": "cr:Field",
7        "dataType": "sc:ImageObject",
8        "source": {
9          "fileSet":{"@id": "image-files"},
10         "extract":{"fileProperty":"content"}
11       }
12     },
13     {
14       "@id": "images/hash",
15       "@type": "cr:Field",
16       "dataType": "sc:Text",
17       "source": {
18         "fileSet": {"@id": "image-files"},
19         "extract": {"fileProperty":
        "filename"},
20         "transform": {"regex":
        "([^\\/]*)\\.jpg"}
21       },
22       "references": {
23         "fileObject": {"@id": "metadata"},
24         "column": "hash"
25       }
26     },
27     { "@id": "images/coordinates",
28       "@type": "cr:Field",
29       "dataType": "sc:GeoCoordinates",
30       "subField": [
31        {   "@id": "images/coordinates/latitude",
32            "@type": "cr:Field",
33            "source": {
34              "fileObject": {"@id": "metadata"},
35              "column": "latitude"}
36        },
37        { "@id": "images/coordinates/longitude",
38         "@type": "cr:Field",
39         "source": {
40              "fileObject": {"@id": "metadata"},
41              "column": "longitude"}
42        }]
43     }]
44 }
```

Figure 4: A RecordSet that joins images and structured metadata from the PASS dataset.

2.3.3 The Structure Layer

While FileObject and FileSet describe a dataset's resources, they lack information on how the content of the resources is organized. This is addressed with RecordSet, which allows loading data of various formats into a standard representation, including structured (CSV and JSON) and unstructured (text, audio, and video) data. Handling all data formatting information in one-layer abstracts away format heterogeneity, addressing a key challenge in processing and loading ML data. RecordSet provides a common structure description for records that may contain multiple fields, which can be used across different modalities. As an example, Figure 4 shows a RecordSet combining images from PASS with additional features from a metadata CSV file. Each Field in the RecordSet defines the source of its data, which may refer to the contents of elements in a FileSet. For instance, the Field images/image_content in line 9 refers to the image-files FileSet and also points to the specific property to extract in line 10.

Fields can be nested, as we can see in the images/coordinates field, which contains two subfields: images/coordinates/latitude and images/coordinates/longitude. Croissant supports nesting entire RecordSets, e.g., to add annotations (e.g., object bounding boxes) to images, where each image may correspond to multiple structured annotations. See Croissant's COCO [30] definition[4] for a representative example. RecordSet also supports joining heterogeneous data and data manipulation methods, like JSON Path and regular expressions, for flexible data extraction and transformation.

2.3.4 The Semantic Layer

The semantic layer introduces a number of useful features in the context of ML data. These are implemented using the primitives defined in the previous sections, generally as new classes or properties defined in the Croissant namespace. Semantic typing is used to describe important aspects of ML practice, such as the dataset splits (train, test, validation) as well as dataset labels. Additionally, semantic typing is used to describe commonly used data types, such as bounding boxes, categorical data, or segmentation masks. As an example, in Figure 4, the structured Field images/coordinates has the dataType GeoCoordinates[5] from schema.org. The subFields images/coordinates/latitude and images/coordinates/longitude are implicitly mapped to the latitude and longitude properties associated with that class, because their names match by suffix.

2.3.5 The Croissant-RAI Extension

Croissant-RAI [7] is an extension of the Croissant format that builds on existing responsible AI (RAI) dataset documentation approaches, such as Data Cards [22] and Datasheets for Datasets [10], making it easier to publish, discover, and reuse RAI metadata. The extension was developed around RAI use cases such as documenting the data life cycle, data labeling and participatory processes, information for AI safety, fairness assessments, and regulatory compliance. It was developed through a multi-step, iterative vocabulary engineering process. Based on the target use cases, a list of properties was defined through evaluation of related dataset documentation vocabularies and the Croissant vocabulary with an aim to detect overlaps and gaps. The resulting properties were evaluated by annotating example datasets to verify their usability and usefulness. For more details, see [31].

2.3.6 Croissant Tools and Integrations

In parallel with the definition of the Croissant format, we have pursued a number of integrations, with the goals of 1) making Croissant immediately useful to users, and 2) grounding Croissant in the requirements of real-world datasets and tools. Figure 1 gives an overview of the Croissant ecosystem.

Data Repositories

Croissant has been integrated into three major dataset repositories: Hugging Face Datasets, Kaggle Datasets, and OpenML, which together describe over 400,000 datasets in the Croissant format. This integration has succeeded with minimal effort because Croissant is an extension of the widely adopted Schema.org/Dataset vocabulary and does not require changing the existing data layout. Supporting Croissant involved adding additional fields to existing metadata. Furthermore, most repositories offer normalized data representations (Hugging Face and OpenML convert most datasets to Parquet) and

[4] https://github.com/mlcommons/croissant/blob/main/datasets/1.0/coco2014/metadata.json

[5] http://schema.org/GeoCoordinates

their own data types (such as relational schemas for tabular data). Consequently, the conversion to Croissant primarily focuses on managing these data formats and specifying associated data types as RecordSet definitions.

In addition to the support from individual data repositories, Croissant is also supported by Google Dataset Search [32]. When a user searches for a query that returns Croissant datasets, a special filter allows them to restrict the results to only Croissant datasets. This functionality allows users to effectively search for Croissant datasets across data repositories and the entire web.

Dataverse implementation

Croissant support was implemented by IQSS (https://www.iq.harvard.edu/) in the Dataverse data repository and available as part of standard distribution as an external metadata exporter[6] in versions higher than 6.2. The current implementation displays a Croissant button on the datasets landing page, allowing users to download them manually or request them via URL with Croissant serialization as a parameter. Once enabled, Croissant export also becomes available in the JSON-LD section of the dataset, replacing the standard schema.org metadata serialization.

The consortium partner DANS[7] has added an experimental Croissant transformation in pyDataverse module based on semantic mappings[8] which can be easily integrated as a part of API microservice or data processing pipeline. This implementation is more flexible and can be reused to add Croissant support for other repositories such as Zenodo and DSpace already providing JSON and OAI-PMH metadata export.

The MuseIT team is using Croissant support as a bridge solution to connect metadata with multimodal content deposited in the MuseIT Dataverse, integrating it with Open Source LLM models such as LLaMa[9] and Mistral[10]. In a nutshell, Croissant has become a universal language for Machine Learning models to communicate with each other and transmit data in a structured format.

ML Frameworks

Croissant's reference implementation is a standalone Python library that supports the validation of Croissant dataset descriptions, their programmatic creation and manipulation, and serialization into JSON-LD. To consume data, the library provides an iterator abstraction that interoperates with existing data loaders. The TensorFlow Datasets [33] library provides a dataset builder[11] that prepares the dataset on disk in a format compatible with JAX, TensorFlow, and PyTorch loaders. Alternatively, frameworks such as PyTorch DataPipes [34] interface with the Croissant library by wrapping the iterator directly. We anticipate that additional optimization opportunities will arise with more varied and larger datasets, perhaps requiring distributed execution as well as more advanced operator scheduling.

Croissant Editor

[6] https://github.com/gdcc/exporter-croissant

[7] https://dans.knaw.nl/en/

[8] https://github.com/Dans-labs/pyDataverse/tree/semantic-mappings

[9] https://www.llama.com

[10] https://docs.mistral.ai/getting-started/models/models_overview /

[11] https://www.tensorflow.org/datasets/format_specific_dataset_builders#croissantbuilder

Croissant is primarily a machine-readable format (in JSON-LD), so users may find it hard to create dataset descriptions by hand. We developed the Croissant Editor[12], (also on GitHub[13]), a tool that lets users visually create and modify Croissant datasets. The Croissant Editor provides form-based editing and validation of Croissant metadata, and bootstraps the definition of resources and RecordSets by inferring them from the data uploaded by the user. The editor integrates the Croissant Responsible AI extension and guides users in describing RAI aspects of their datasets.

2.4. Croissant Evaluation: A User Study

This section describes the user study we conducted to evaluate the Croissant metadata format. We asked ML practitioners to annotate a variety of datasets commonly used in the ML community. Human annotators authored a subset of the Croissant and Croissant-RAI attributes and assessed them based on criteria commonly used in vocabulary evaluation [35].

2.4.1. The User Study Process

Table 1: Post-annotation assessment: Criteria, corresponding questions, and answer scales.

Criteria	**Question**	**Answer Options**
Answer Confidence	How confident are you that your provided annotations are correct?	1 (no confidence) - 5 (very confident that annotations are correct)
Dataset Understanding	How well did you understand the dataset (e.g., the task, domain, modality, etc.)?	1 (I don't understand the dataset at all) - 5 (the dataset incl. its purpose, creation, etc. is very clear and understandable for me)
Completeness	Is there any (in your opinion important) information about the dataset which you can't define using Croissant?	1 (yes, there is lots of critical information about the dataset that Croissant does not capture) - 5 (no, every important information about this dataset, which might be useful for ML users, is captured in Croissant attributes)
Conciseness	Did you find any attributes redundant and not definable for this dataset?	1 (yes, there are lots of redundant attributes) - 5 (no, none of the attributes is redundant)
Readability	How intuitive are the attributes names for you? A name is not intuitive if you need to check the specification to understand the attribute's name?	1 (not intuitive at all, for each single attribute I checked the specification to understand it) - 5 (very intuitive, based on the name I could understand the attribute very well)
Understandability	Rate the ease of understanding the Croissant specification.	1 (Understanding the spec. was very hard) - 5 (the spec. is very easy to understand)

Recruitment of Annotators and Annotation Process

We recruited nine volunteers from the Croissant development community, all proficient in English with backgrounds in vocabulary and ontology engineering, dataset documentation, and responsible AI. We collected demographic information from all annotators, which we published in the user study report

[12] https://huggingface.co/spaces/MLCommons/croissant-editor

[13] https://github.com/mlcommons/croissant/tree/main/editor

[36]. For each dataset, we collected metadata definitions from three annotators, resulting in thirty annotations. Each annotator assessed approximately three datasets on average, with three annotating one dataset and one annotating six datasets.

The instructions for the annotators included: (i) a brief introduction to the Croissant metadata format, (ii) the purpose of the user study, (iii) the definitions of the requested Croissant attributes, (iv) links to format specifications, and (v) a link to each dataset. Prior to starting the study, we obtained ethical clearance and informed the annotators about the data being collected. For each dataset, annotators filled out a provided template with the attributes to complete, after which they answered questions about their understanding of the datasets and their confidence in the annotations on a Likert scale [37].

Selection of Croissant Attributes

We selected a set of attributes from Croissant's Dataset Layer and the Croissant-RAI attributes. These attributes were chosen because they (i) require manual specification, (ii) can be defined by users using the dataset itself or a publication describing the dataset, and (iii) support the discoverability and reproducibility of datasets. For example, missing or limited descriptions of datasets reduce their discoverability and hinder their use [38]. We also selected attributes that would ensure datasets could be reproduced in the same conditions as intended.

Table 2: Annotated Croissant attributes.

Property
sc:description
sc:license
sc:name
sc:url
sc:creator
sc:publisher
sc:datePublished
sc:inLanguage
cr:citeAs
cr:isLiveDataset

Table 3: Annotated Croissant-RAI attributes.

Property	**RAI Use Case**
rai:dataCollection	Data life cycle
rai:dataCollectionTimeframe	Data life cycle
rai:dataAnnotationPlatform	Data labelling
rai:annotatorDemographics	Data labelling
rai:dataUseCases	AI safety and fairness evaluation
rai:personalSensitiveInformation	Compliance

Table 4: Annotated datasets.

Dataset	Modality
MMLU	Language
Dolly-15k	Language
FLORES	Language
CIFAR10	Vision
MSCOCO	Vision
Visual Genome	Vision
MMMU	VL
MathVista	VL
MLS_Eng	Audio
librispeech_asr	Audio

ML Datasets

We selected commonly used datasets from the language, vision, and audio modalities based on their popularity and availability in repositories like Hugging Face. These datasets had pre-existing descriptions and were associated with publications describing their creation (Table 4).

Evaluation

We evaluated the collected attribute annotations by assessing the agreement among annotators. For textual attributes, we measured similarity between attribute annotations using a BLEU (Bilingual Evaluation Understudy) score, which indicates how well the annotations matched in terms of text similarity [39].

2.4.2. Mapping Evaluation Criteria to Croissant

We evaluated Croissant based on five key criteria commonly used in vocabulary evaluation [35, 40] (evaluating vocabularies based on consistency, completeness, conciseness, readability, and understandability is common practice in the field of dataset documentation and metadata):

1. **Consistency.** We assessed how well annotations for the same attribute and dataset aligned, indicating the consistency of the vocabulary.
2. **Completeness.** We evaluated whether Croissant covers all necessary attributes to capture important information about datasets. We also asked annotators to flag any missing information that could not be defined using Croissant.

3. **Conciseness.** We asked annotators whether they found any attributes redundant or not definable for the datasets. This helped assess whether the vocabulary avoided unnecessary definitions.
4. **Readability.** We evaluated how intuitive the attribute names were. Annotators were asked to rate the attribute names for their clarity and ease of understanding.
5. **Understandability.** We evaluated how easily annotators could understand the attributes from the provided documentation. We instructed annotators to use the Croissant specifications [7, 9] (croissant specifications were used as a reference point for understanding and interpreting the metadata attributes during the study) and prompted them with questions afterward.

2.4.3. Results and Discussion

This section discusses the findings from the user study.

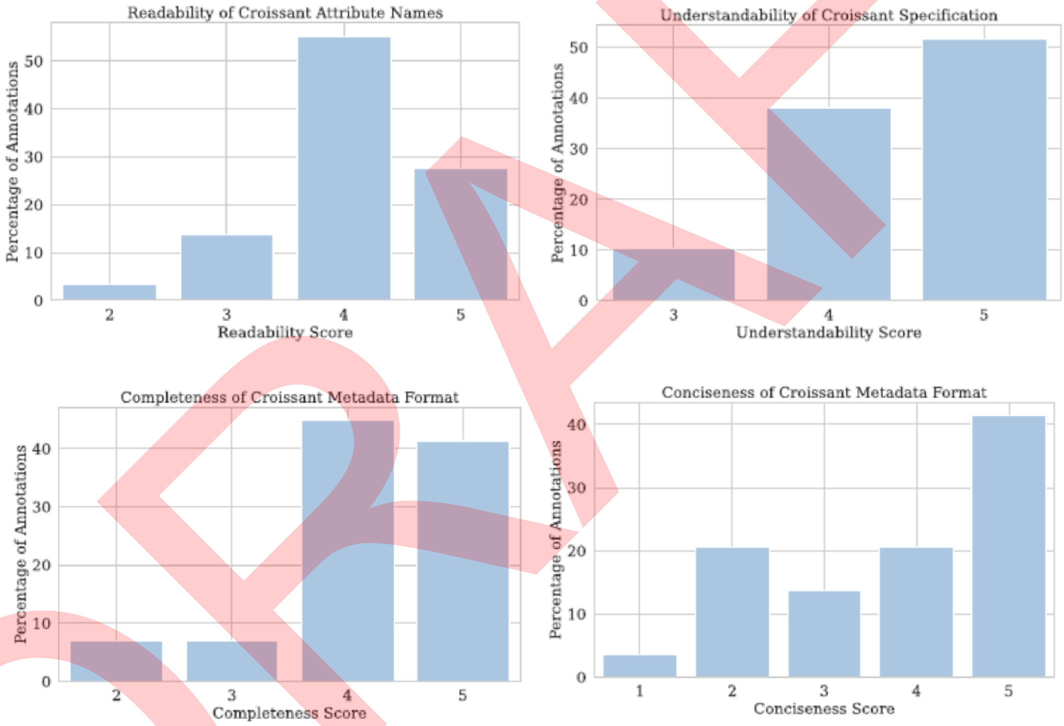


Figure 5: Answers to the questions on Readability, Understability, Completeness and Conciseness

Criteria Evaluation. Over 80% of the annotations indicated that Croissant attributes captured important information about the datasets (Figure 5). For the conciseness criterion, some annotators found a few attributes redundant or difficult to define, especially with some of the Croissant-RAI attributes, which required additional context that was not available in the dataset documentation. However, the majority of annotators found the attribute names intuitive, resulting in high readability scores. Most annotators also found the specifications understandable, which gave us confidence in the data collected during the study.

In addition to the criteria-related questions, we asked annotators about their confidence in the correctness of their annotations and their understanding of the datasets. The majority of annotators expressed high confidence in their annotations, with more than 75% selecting a high confidence rating.

Attributes Evaluation. We assessed the agreement among annotators using BLEU scores (Table 5). Overall, the average BLEU score for Croissant attributes was higher than for Croissant-RAI attributes. This difference can be attributed to the fact that some Croissant-RAI attributes require free-form text answers, which can vary more across annotations. Croissant attributes, on the other hand, are more easily extractable from dataset documentation and often involve predefined values like language or license type, which led to higher agreement.

Table 5: BLEU scores for annotated datasets and attributes (i.e. description, license, url, creator, publisher, datePublished, inLanguage, citeAs, dataCollection, dataCollectionTimeframe, dataAnnotationPlatform, annotatorDemographics, dataUseCases, personalSensitiveInformation)

Dataset	desc	lic	url	creator	publ	datePub	lang	citeAs	dataCol	time	plat	demogr	useCases	persInfo
flores	0.03	0.6	0.45	0.88	0.54	0.12	0.84	0.31	0.4	0.08	0.34	0.0	0.42	0.0
cifar-10	0.39	1.0	0.31	0.17	0.16	0.14	1.0	0.26	0.35	0.0	1.0	0.0	0.29	1.0
dolly-15k	0.56	1.0	1.0	0.82	0.5	0.28	0.34	0.75	0.57	0.0	1.0	0.0	0.39	0.01
mscoco	0.7	1.0	0.65	0.26	0.0	0.24	1.0	0.0	0.32	1.0	0.78	0.0	0.88	0.0
visual gen	0.41	1.0	0.18	0.49	0.0	0.51	1.0	1.0	0.29	0.19	0.0	0.84	0.27	1.0
mmmu	0.89	0.49	1.0	0.76	0.33	0.21	1.0	1.0	0.77	1.0	0.0	0.05	0.48	0.62
mmlu	0.13	0.0	0.56	0.97	0.37	0.32	1.0	0.79	0.6	1.0	0.07	0.65	0.45	0.0
mathvista	1.0	0.34	0.57	0.53	0.07	0.26	0.13	1.0	0.16	1.0	0.0	0.05	0.22	1.0
mls_eng	0.35	1.0	1.0	0.64	0.35	0.56	0.03	1.0	0.3	1.0	0.0	1.0	0.36	0.0
librispeech	0.73	1.0	0.17	0.82	0.33	0.44	1.0	0.04	0.34	1.0	0.0	0.29	0.25	0.21
Average	0.52	0.74	0.59	0.63	0.26	0.31	0.73	0.62	0.41	0.63	0.32	0.29	0.4	0.38
Median	0.52	1.0	0.57	0.64	0.33	0.28	1.0	0.75	0.35	1.0	0.07	0.05	0.39	0.21

2.5. Limitations and Future Work

The Croissant metadata format provides a shared representation across various tools and platforms for managing digital cultural heritage assets. While it facilitates consistent metadata documentation, certain challenges remain that should be addressed in future work. First, its structure may pose difficulties for users unfamiliar with the format, which could hinder its adoption across different domains. To improve accessibility and usability, we plan to extend Croissant tools (such as the Croissant editor) and provide comprehensive documentation, which are essential for making Croissant datasets easier for users to utilize. This includes adding annotated dataset examples in the Croissant repository and developing community guidelines that take into account domain-specific needs.

Second, the Croissant editor, as an interface for creating metadata, is still in its early stages and will be enhanced in future work. Upcoming improvements will include support for functionalities such as file archiving, nested fields, and additional metadata features. Finally, to further demonstrate Croissant's versatility in handling complex data, we plan to release a GeoSpatial extension for Croissant (Geo-Croissant), which will support handling of specialized file formats like HDF5 and Zarr, with accompanying examples.

2.6. Conclusions

This section introduced Croissant, a metadata format designed to improve the management and sharing of digital cultural heritage assets. Croissant enhances the discoverability, portability, and interoperability of datasets across a variety of platforms, repositories, and tools, addressing key challenges in the preservation and accessibility of cultural heritage data. By providing a standardized data representation, Croissant ensures that cultural heritage datasets can be easily shared, preserved, and utilized across different systems.

Croissant has already been adopted by several prominent platforms, and its metadata has been positively evaluated for being readable, understandable, complete, and concise by human raters. However, for Croissant to fully realize its potential in the cultural heritage sector, wider adoption within both cultural heritage institutions and related industries is crucial. The availability of more Croissant datasets and support from relevant tools and platforms will be key to its success. We encourage institutions, researchers, and developers to join the Croissant community and contribute to its growth.

Finally, the extensible nature of Croissant, along with its ability to represent diverse types of cultural heritage data, offers a unique opportunity for communities to adapt the format for their specific needs. Extensions like the Croissant-RAI for Responsible AI highlight Croissant's potential to facilitate cross-disciplinary collaboration, particularly between the cultural heritage and technology sectors, ensuring that digital cultural assets are managed and shared ethically and responsibly.

3. IPR Management through Data Trust and Provenance Technologies: Insights from Multimodal Fact-Checking

3.1 Introduction

As digital cultural heritage assets are increasingly shared and transformed into digital representations, ensuring IPR protection and authenticity verification becomes essential. Digital Rights Management (DRM) mechanisms must evolve beyond traditional text-based metadata verification to address the complexities of multimodal digital content. Images, videos, and textual descriptions can be altered, misrepresented, or used out of context, posing significant risks to cultural preservation efforts.

To address these challenges, MuseIT integrates structured metadata documentation with automated verification techniques. Automated verification methods designed for multimodal fact-checking provide a robust framework for tracking the integrity of digital artifacts, ensuring their correct attribution, and preventing unauthorized modifications. While Croissant ensures clear documentation of dataset provenance, licensing, and structure, metadata alone cannot fully safeguard cultural heritage content from manipulation or misattribution. Therefore, MuseIT incorporates insights from multimodal automated fact-checking, a verification framework that systematically verifies the authenticity of content by analyzing text, images, audio, and video in combination.

Multimodal Automated Fact-Checking provides a structured approach to addressing these challenges. By leveraging methodologies designed for detecting misinformation and manipulated media, we can enhance metadata integrity, traceability, and access control for digital heritage assets. In our survey paper [2] titled "Multimodal Automated Fact-Checking: A Survey" published in the Findings of EMNLP 2023, we have outlined a conceptual framework for multimodal verification, detailing how different modalities—text, image, audio, and video—can be systematically analyzed for fact-checking in line with a DRM framework for a robust strategy for IPR protection.

By combining metadata documentation with AFC methods, MuseIT adopts a dual-layered approach to IPR protection. Croissant ensures comprehensive metadata coverage, while AFC techniques provide evidence retrieval, manipulation detection, and provenance tracking capabilities. Together, these approaches strengthen MuseIT's ability to manage digital cultural heritage assets securely and responsibly.

3.2. Multimodal Fact-Checking as a Verification Framework

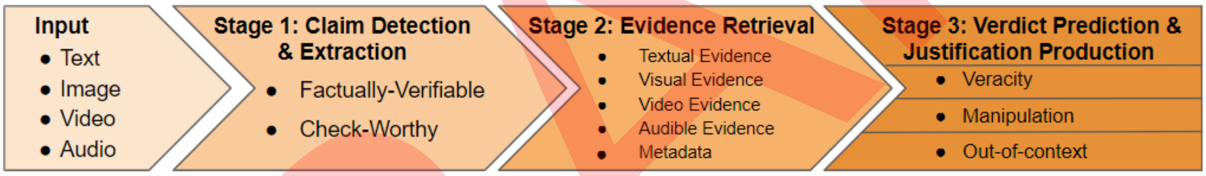


Figure 6: Multimodal Fact-Checking Pipeline

Figure 6 illustrates the multimodal fact-checking pipeline, highlighting the sequential stages of claim detection and extraction, evidence retrieval, and verdict prediction, with specific emphasis on handling different modalities such as text, images, audio, and video to ensure comprehensive verification.

Misinformation and misrepresentation in digital heritage archives often involve multiple modalities, including images with misleading captions, manipulated videos, or falsified textual claims. The nature of multimodal content requires verification systems capable of analyzing interconnections between different forms of media to detect inconsistencies and ensure the authenticity of digital assets.

Multimodal content is particularly susceptible to misinterpretation and deliberate manipulation, as misinformation is often conveyed in multiple modalities, e.g., a miscaptioned image. Multimodal misinformation is perceived as more credible by humans and spreads faster than its text-only counterparts. Ensuring metadata integrity and provenance tracking across all modalities is critical for maintaining the reliability of digital cultural heritage representations.

3.2.1. Claim Detection and Metadata Integrity

A primary challenge in managing IPR and digital authenticity is identifying claims related to ownership, authenticity, and historical accuracy. In many cases, claims about digital artifacts are embedded within images, videos, or accompanying text, making it necessary to extract and verify these claims before validating their authenticity.

The process of claim detection focuses on identifying factually verifiable and contextually significant statements within multimodal content:

The first pipeline stage aims to find checkable (i.e., factually verifiable) and check-worthy (i.e., important factual) claims. Multimodal claims can be diverse and include: (1) a written claim embedded in another modality, such as an image or a spoken claim in an audio or video; (2) a claim that a piece of content is authentic, e.g., that a video footage is from a specific geographic location; (3) a claim for which the evidence is manipulated to support it, e.g., through lip-syncing.

Ensuring metadata integrity requires automated claim extraction across text, image, and audio modalities. Technologies such as OCR (Optical Character Recognition) for image-based text, speech-to-text transcription for audio claims, and metadata comparison tools can help identify discrepancies in authorship, timestamps, or licensing details. By integrating these mechanisms into digital rights management systems, it becomes possible to track, authenticate, and verify ownership claims, mitigating the risks of misattribution or misuse.

3.2.2. Evidence Retrieval and Provenance Tracking

Once claims are identified, verifying their authenticity requires retrieving supporting evidence. Provenance tracking involves ensuring that digital artifacts maintain a verifiable history of their creation, ownership, and modifications. The process of evidence retrieval can take two forms:

Similarly to fact-checking with text, multimodal fact-checking often relies on evidence to make judgments, similar to the process followed by human fact-checkers. Two main approaches have been used in the past: (i) using the claim to be checked as evidence itself, e.g., to detect manipulation; and (ii) retrieving additional evidence.

For digital cultural heritage, provenance verification can be strengthened by integrating institutional records, blockchain-based authentication, and multimodal retrieval methods. If an artifact is digitized as a 3D model, image, or textual description, verification mechanisms should cross-reference it with registered metadata, archival databases, and linked open data repositories to confirm its historical accuracy and rightful attribution.

Additionally, multimodal retrieval methods can be used to identify unauthorized modifications or reproductions. If a manipulated version of a historical image circulates online, automated systems should be able to trace the original source, compare it against registered copies, and detect alterations in color, composition, or contextual captions.

3.2.3. Manipulation Detection and Content Integrity

A major risk in digital archives is the deliberate or unintentional alteration of cultural heritage artifacts, where images, videos, or audio recordings are edited to misrepresent historical content. Identifying these modifications requires a robust detection framework capable of analyzing manipulated content across different media types.

Manipulation classification commonly addresses (i) misinformative claims with manipulated content; (ii) correct claims accompanied by manipulated content (e.g., to increase credibility). Many methods exist to manipulate text, visual, and audio content. While some require more knowledge to use (e.g., speech synthesis), other manipulations can be achieved with simple tools (e.g., changing speed of videos).

To ensure content integrity, detection methods should include image forensics, deepfake detection models, and watermark validation. Neural networks such as Vision Transformers (ViTs), recurrent convolutional networks (RCNs), and graph-based multimodal encoders can be leveraged to detect inconsistencies in visual, textual, and auditory data. These systems help ensure that historical artifacts remain unaltered and faithfully represented in digital repositories.

Out-of-Context Detection and Contextual Authenticity

Beyond direct manipulation, another significant challenge is the misuse of unchanged content in misleading contexts. This problem is particularly relevant for historical images or videos that are repurposed to support false narratives.

Using unchanged content out-of-context is one of the most common and easiest methods to create multimodal misinformation. Recent work has also studied the applicability of traditional multimodal misinformation detection methods to identify out-of-context content.

To address this, verification mechanisms should monitor how digital assets are being presented across online platforms. If a museum artifact is used in a misleading political or historical claim, automated tracking can detect its improper usage and verify it against metadata records. Context-aware verification can also compare images, text, and video descriptions to flag potential misrepresentations.

3.3. Challenges and Future Directions

3.3.1. Claim extraction from multimodal content

Multimodal claims, e.g., manipulated videos, are often embedded in specific contexts and framed as (part of) larger stories. For example, countering the misinformation in images requires not only classifying if the image is manipulated but understanding the context of the depiction being shown in the image as well. Only then can relevant evidence data be extracted and used to verify the claims of the image. Determining what is being claimed is a challenging first step in multimodal automated fact-checking. However, current efforts for multimodal claim extraction are limited to text extraction from visual content or transcribing audios and videos [41, 42, 43]. Addressing this challenge will require modeling approaches to effectively align and integrate all modalities present in and around the claim. For example, methods for pixel-based language modeling have recently been introduced to better align visually situated language with image content [44]. Such approaches considering modalities beyond text and vision for multimodal data alignment can be useful for claim extracting from multimodal input.

3.3.2. Multimodal evidence retrieval

Evidence retrieval for audio and video fact-checking remains a major challenge. Different from other modalities, they cannot be easily searched on the web or social media networks [45]. Fact-checkers often use text accompanying the videos to find evidence [45]. Reverse image search engines, e.g., Google Lens or TinEye, require screenshots from the video as input – and thus require the correct timeframe, which can be challenging to extract. A dedicated adversary can render current tools very difficult to use. Very often, evidence for image or audio fact-checking is retrieved using text accompanying them, e.g., metadata, social media comments, or captions [46, 47, 48, 49]. While incorporating the textual information and the other modality (e.g., audio/image) in retrieval would provide more information, this is currently missing. How to best retrieve evidence data that is non-textual or has a different modality than the claim remains a challenge.

3.3.3. Generalizing detection of visual manipulations

The recent popularity of diffusion models (DMs) for visual manipulation has raised questions regarding the generalizability of manipulation detectors developed for earlier models (e.g., GANs [40]). Detection models are biased towards specific manipulation models and struggle to generalize [50, 51]. A recent study [51] shows that detectors initially developed for GANs have average performance drops of around 15% for images by DMs. While new detection approaches for DM manipulations are already being developed [52, 53], the question of how to generalize and increase the robustness of manipulation detectors for potential future manipulation models remains open. Potential solutions can include evidence-based approaches, where the manipulated content is used to retrieve evidence data (e.g., the original video or counterfactual evidence) to prove the manipulation.

3.3.4. Justifications for multimodal fact-checking

While explainable fact-checking has received attention recently [54, 55], there is limited work on producing justifications for multimodal content. Previous efforts on multimodal justification production have mostly focused on highlighting parts of the input to increase interpretability [56, 57]. Natural language justifications that explain the fact-check of multimodal claims so that they are accessible to non-technical audiences have not yet been developed. To develop solutions, we first need appropriate benchmarks to measure progress. Moreover, with the recent advances of neural models for visual and audio generation and editing, another so far unexplored direction presents itself: editing input images/videos/audios or generating entirely new content to explain fact-checking results. This could include, for example, the generation of infographics or video clips to explain fact-checks. Such a system, especially if guided by human fact-checkers [58], would be a potent tool. As noted in [59], “well-designed graphs, videos, photos, and other semantic aids can be helpful to convey corrections involving complex or statistical information clearly and concisely.”

3.4. Conclusion

Ensuring IPR compliance and authenticity verification in digital cultural heritage requires advanced provenance tracking, manipulation detection, and metadata integrity analysis. A structured multimodal verification framework allows for cross-modal authentication of text, images, and videos, ensuring that digital assets remain protected against unauthorized modifications or misattribution.

By integrating automated claim detection, multimodal evidence retrieval, and manipulation classification techniques, digital rights management systems can provide verifiable authenticity records, prevent unauthorized reuse, and maintain the historical integrity of digital artifacts. This approach safeguards cultural assets against distortion, ensuring that future generations have access to accurate and trustworthy representations of heritage materials.

4. Overall Conclusions

This report has examined the critical challenges involved in managing, documenting, and verifying multimodal cultural heritage assets in the digital realm. It introduced **Croissant**, a metadata format designed for ML-ready datasets, providing a structured approach to enhance dataset discoverability, interoperability, and responsible AI integration. By offering standardized descriptions of multimodal content, Croissant facilitates improved provenance tracking, licensing transparency, and IPR compliance. Additionally, the report explored **multimodal fact-checking**, a robust mechanism for content authenticity verification, incorporating techniques such as claim detection, evidence retrieval, and manipulation detection. Together, these approaches form a comprehensive framework for safeguarding the integrity and responsible use of digital assets.

The integration of Croissant and multimodal fact-checking directly aligns with the goals of the **MuseIT project**, which focuses on the preservation and digitization of cultural heritage. By combining Croissant's structured metadata framework with automated verification through multimodal fact-checking, MuseIT ensures the seamless management, authentication, and protection of digital assets. Croissant's ability to organize metadata and track provenance complements MuseIT's mission by enabling a transparent and legally compliant ecosystem for cultural artifacts. The addition of multimodal fact-checking enhances this further, strengthening MuseIT's capability to prevent misinformation, validate historical accuracy, and preserve the authenticity of digitized heritage content.

Looking ahead, future work will aim to extend Croissant's capabilities to support increasingly complex multimodal datasets, ensuring broader adoption across repositories and ML frameworks. In parallel, continued refinement of multimodal fact-checking techniques will be essential to enhance their resilience against advanced content manipulations. Finally, collaborations with cultural institutions and policymakers will be crucial for aligning these frameworks with evolving regulatory and ethical standards. Through ongoing development, the MuseIT project will continue to advance as a sustainable, trustworthy, and scalable solution for managing and protecting digital cultural heritage in the digital age.

Acknowledgements

The research documented in this deliverable bases on the grant of the HE project MuseIT, which has been co-founded by the European Union under the Grant Agreement No 101061441. Views and opinions expressed are, however, those of the authors and do not necessarily reflect those of the European Union or European Research Executive Agency. We thank the reviewers Roberto Maffulli, Andrea Scharnhorst for their insightful feedback and suggestions for improving this report.

References

1. Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Joan Giner-Miguelez, Pieter Gijsbers, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Michael Kuchnik, Satyapriya Krishna, Sylvain Lesage, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Hamidah Oderinwale, Pierre Ruyssen, Tim Santos, Rajat Shinde, Elena Simperl, Arjun Suresh, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Susheel Varma, Jos van der Velde, Steffen Vogler, Carole-Jean Wu, and Luyao Zhang. "Croissant: A metadata format for ml-ready datasets." In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 82133–82148. Curran Associates, Inc., 2024.
2. Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal Automated Fact-Checking: A Survey. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5430–5448, Singapore. Association for Computational Linguistics.
3. Daga, Enrico, Jason Carvalho, and Alba Morales Tirado. Extracting licence information from web resources with a Large Language Model. (2024).
4. Daga, E., Carvalho, J., Gurrieri, M. and Scharnhorst, A., 2024. D2. 6: Ontology of licencing, ownership and conditions of use (V1. 0).Michael Kuchnik, Ana Klimovic, Jiri Simsa, Virginia Smith, and George Amvrosiadis. Plumber: Diagnosing and removing performance bottlenecks in machine learning data pipelines. Proceedings of Machine Learning and Systems, 4:33–51, 2022.
5. Michael Kuchnik, Ana Klimovic, Jiri Simsa, Virginia Smith, and George Amvrosiadis. Plumber: Diagnosing and removing performance bottlenecks in machine learning data pipelines. Proceedings of Machine Learning and Systems, 4:33–51, 2022.
6. Luis Oala, Manil Maskey, Lilith Bat-Leah, Alicia Parrish, Nezihe Merve Gürel, Tzu-Sheng Kuo, Yang Liu, Rotem Dror, Danilo Brajovic, Xiaozhe Yao, Max Bartolo, William A Gaviria Rojas, Ryan Hileman, Rainier Aliment, Michael W. Mahoney, Meg Risdal, Matthew Lease, Wojciech Samek, Debojyoti Dutta, Curtis G Northcutt, Cody Coleman, Braden Hancock, Bernard Koch, Girmaw Abebe Tadesse, Bojan Karla, Ahmed Alaa, Adji Bousso Dieng, Natasha Noy, Vijay Janapa Reddi, James Zou, Praveen Paritosh, Mihaela van der Schaar, Kurt Bollacker, Lora Aroyo, Ce Zhang, Joaquin Vanschoren, Isabelle Guyon, and Peter Mattson. DMLR: Data-centric machine learning research - past, present and future. Journal of Datacentric Machine Learning Research, 2024. URL https://openreview.net/forum?id=2kpu78QdeE. Featured Certification, Survey Certification.
7. Mubashara Akhtar, Nitisha Jain, Joan Giner-Miguelez, Omar Benjelloun, Elena Simperl, Lora Aroyo, Rajat Shinde, Luis Oala, and Michael Kuchnik. Croissant RAI Specification. Technical report, 2024. URL https://mlcommons.org/croissant/RAI/1.0.
8. Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2021.
9. Omar Benjelloun, Elena Simperl, Pierre Marcenac, Pierre Ruyssen, Costanza Conforti, Michael Kuchnik, Jos van der Velde, Luis Oala, Steffen Vogler, Mubashara Akhtar, Nitisha Jain, and Slava Tykhonov. Croissant format specification. Technical report, 2024. URL https://mlcommons.org/croissant/1.0.
10. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2021.
11. RSI Wilson, JS Goonetillake, WA Indika, and Athula Ginige. A conceptual model for ontology quality assessment. Semantic Web, (Preprint):1–47, 2022.

12. Riccardo Albertoni, David Browning, Simon J D Cox, Alejandra Gonzalez Beltran, Andrea Perego, and Peter Winstanley. Data catalog vocabulary (DCAT) - version 3. https://www.w3.org/TR/vocab-dcat-3/, 01 2024. (Accessed on 03/18/2024).
13. schema.org. Schema.org v26.0. https://github.com/schemaorg/schemaorg/tree/main/data/releases/26.0/, 02 2024. (Accessed on 03/18/2024).
14. Ramanathan V Guha, Dan Brickley, and Steve Macbeth. Schema. org: evolution of structured data on the web. Communications of the ACM, 59(2):44–51, 2016.
15. Frictionless Working Group. Data packages. https://specs.frictionlessdata.io/, 2024. (Accessed on 03/21/2024).
16. W3C Working Group. CSV on the web: A primer. https://www.w3.org/TR/tabular-data-primer/, 2016. (Accessed on 03/21/2024).
17. Jonathan Lawson, Moran N Cabili, Giselle Kerry, Tiffany Boughtwood, Adrian Thorogood, Pinar Alper, Sarion R Bowers, Rebecca R Boyles, Anthony J Brookes, Matthew Brush, et al. The data use ontology to streamline responsible access to human biomedical datasets. Cell Genomics, 1(2), 2021.
18. Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. Scientific data, 3(1):1–9, 2016.
19. Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. The principles of data-centric ai. Commun. ACM, 66(8):84–92, jul 2023. ISSN 0001-0782. doi: 10.1145/3571724. URL https://doi.org/10.1145/3571724.
20. Nathalie A. Smuha. The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence. Computer Law Review International, 20(4):97–106, 2019. doi: doi:10.9785/cri-2019-200402. URL https://doi.org/10.9785/cri-2019-200402.
21. Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.
22. Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai, 2022.
23. Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards, 2018.
24. Kaggle. Kaggle datasets: A platform for data science competitions and collaborative work, 2024. URL https://www.kaggle.com/datasets.
25. Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. SIGKDD Explorations, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL http://doi.acm.org/10.1145/2641190.264119.
26. HuggingFace. Hugging Face Datasets: A community-driven hub for ready-to-use datasets, 2024. URL https://huggingface.co/datasets.
27. HuggingFace. Hugging Face dataset cards. https://huggingface.co/docs/hub/en/datasets-cards, 2024. (Accessed on 06/05/2024).
28. Mubashara Akhtar, Nitisha Jain, Joan Giner-Miguelez, Omar Benjelloun, Elena Simperl, Lora Aroyo, Rajat Shinde, Luis Oala, and Michael Kuchnik. Croissant RAI Specification. Technical report, 2024. URL https://mlcommons.org/croissant/RAI/1.0.
29. Yuki M. Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. NeurIPS Track on Datasets and Benchmarks, 2021.
30. Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

31. Nitisha Jain, Mubashara Akhtar, Joan Giner-Miguelez, Rajat Shinde, Joaquin Vanschoren, Steffen Vogler, Sujata Goswami, Yuhan Rao, Tim Santos, Luis Oala, Michalis Karamousadakis, Manil Maskey, Pierre Marcenac, Costanza Conforti, Michael Kuchnik, Lora Aroyo, Omar Benjelloun, and Elena Simperl. A Standardized Machine-readable Dataset Documentation Format for Responsible AI. to appear in ArXiv, abs/5643361, 2024.
32. Dan Brickley, Matthew Burgess, and Natasha Noy. Google dataset search: Building a search engine for datasets in an open web ecosystem. In The world wide web conference, pages 1365–1375, 2019.
33. TFDS. TensorFlow Datasets, a collection of ready-to-use datasets. https://www.tensorflow.org/datasets, 03 2024.
34. PyTorch. DataPipe Tutorial. https://pytorch.org/data/beta/dp_tutorial.html, 2024. (Accessed on 10/28/2024).
35. Asunción Gómez-Pérez. Evaluation of ontologies. Int. J. Intell. Syst., 16(3):391–409, 2001.
36. Croissant Working Group. Croissant - user research report, August 2024. URL https://doi.org/10.5281/zenodo.13350974.
37. Rensis Likert. A technique for the measurement of attitudes. Archives of Psychology, 22(140):1–55, 1932.
38. Rebecca M. Deutsch, Jacob G. Foster, Matthew S. Peters, and Guy H. Hembrooke. Maintaining Data Quality in Large-Scale Data-Centric AI Research. Journal of Data Science, 29(4):37–51, 2024.
39. Anna Klopfenstein and Frank Löser. Automatic Generation of Quality Metrics for Databases and Datasets. Journal of Computational Data Management, 12(3):1–21, 2023.
40. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144.
41. Qu, J., Li, L. H., Zhao, J., Dev, S., & Chang, K. W. (2022). Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *CoRR*, abs/2205.12617.
42. Garimella, K., & Eckles, D. (2020). Images and misinformation in political groups: Evidence from WhatsApp in India. *CoRR*, abs/2005.09784.
43. Maros, M., Bourgonje, P., & Rehm, G. (2021). Automatic fact-checking of spoken claims: Introducing the factuality corpus for Dutch. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1016–1026.
44. Lee, K., Joshi, M., Turc, I., Hu, H., Liu, F., Eisenschlos, J., Khandelwal, U., Shaw, P., Chang, M. W., & Toutanova, K. (2022). Pix2struct: Screenshot parsing as pretraining for visual language understanding. *CoRR*, abs/2210.03347.
45. Silverman, C. (2013). *Verification handbook*. European Journalism Centre.
46. Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. *Proceedings of the 22nd International World Wide Web Conference (WWW)*, 729–736.
47. Huh, M., Liu, A., Owens, A., & Efros, A. A. (2018). Fighting fake news: Image splice detection via learned self-consistency. *Proceedings of the European Conference on Computer Vision (ECCV)*, 101–117.
48. Müller-Budack, E., Dörpinghaus, M., & Ewerth, R. (2020). Multimodal misinformation detection: Approaches, challenges, and future directions. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 102–118.
49. Kopev, A., Gusev, I., Burtsev, M., & Filchenkov, A. (2019). Deception detection in multimodal social media data. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1019–1030.
50. Wu, T., Shao, R., & Liu, Z. (2023a). Detecting and grounding multi-modal media manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6904–6913.

51. Ricker, J., Damm, S., Holz, T., & Fischer, A. (2022). Towards the detection of diffusion model deepfakes. *CoRR*, abs/2210.14571.
52. Guarnera, L., Giudice, O., & Battiato, S. (2023). Level up the deepfake detection: A method to effectively discriminate images generated by GAN architectures and diffusion models. *CoRR*, abs/2303.00608.
53. Wu, T., Shao, R., & Liu, Z. (2023b). Detecting and grounding multi-modal media manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6904–6913.
54. Kotonya, N., & Toni, F. (2020b). Explainable automated fact-checking: A survey. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1006–1020.
55. Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). Generating fact checking explanations with explainable fact checking models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7352–7364.
56. Kou, Z., Shang, L., Zhang, Y., & Wang, D. (2020). A multimodal misinformation detector for COVID-19 short videos on TikTok. *Proceedings of the 2021 IEEE International Conference on Big Data (Big Data)*, 899–908.
57. Shang, L., Kou, Z., Zhang, Y., & Wang, D. (2022). A duo-generative approach to explainable multimodal COVID-19 misinformation detection. *Proceedings of the ACM Web Conference (WWW)*, 3623–3631.
58. Nakov, P., Barrón-Cedeño, A., Elsayed, T., Shaar, S., Atanasova, P., & Da San Martino, G. (2021). Automated fact-checking for combating digital misinformation: The role of natural language processing. *Journal of Artificial Intelligence Research (JAIR)*, 134, 177–231.
59. Lewandowsky, S., Cook, J., & Lombardi, D. (2020). *Debunking Handbook 2020*.