This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870811

SPICE

Social cohesion, Participation, and Inclusion through Cultural Engagement

D4.5 Provenance and Process analysis layer: Requirements analysis

Deliverable information	
WP	WP4
Document dissemination level	PU Public
Deliverable type	R Document, report
Lead beneficiary	OU
Contributors	OU, UH
Date	26/04/2022
Document status	Final
Document version	V1.0

INTENTIONALLY BLANK PAGE

Project information

Project start date: 1st of May 2020

Project Duration: 36 months

Project website: https://spice-h2020.eu

Project contacts

Project Coordinator

Silvio Peroni

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

Department of Classical Philology and Italian Studies – FICLIT

E-mail: silvio.peroni@unibo.it

Scientific Coordinator

Aldo Gangemi

Institute for Cognitive Sciences and Technologies of the Italian National Research Council

E-mail: aldo.gangemi@unibo.it

Project Manager

Adriana Dascultu

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

Executive Support Services

E-mail: adriana.dascultu@unibo.it

SPICE consortium

No.	Short name	Institution name	Country
1	UNIBO	ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA	Italy
2	AALTO	AALTO KORKEAKOULUSAATIO SR	Finland
3	DMH	DESIGNMUSEON SAATIO - STIFTELSEN FOR DESIGNMUSEET SR	Finland
4	AAU	AALBORG UNIVERSITET	Denmark
5	OU	THE OPEN UNIVERSITY	United Kingdom
6	IMMA	IRISH MUSEUM OF MODERN ART COMPANY	Ireland
7	GVAM	GVAM GUIAS INTERACTIVAS SL	Spain
8	PG	PADAONE GAMES SL	Spain
9	UCM	UNIVERSIDAD COMPLUTENSE DE MADRID	Spain
10	UNITO	UNIVERSITA DEGLI STUDI DI TORINO	Italy
11	FTM	FONDAZIONE TORINO MUSEI	Italy
12	CELI	CELI SRL	Italy
13	UH	UNIVERSITY OF HAIFA	Israel
14	CNR	CONSIGLIO NAZIONALE DELLE RICERCHE	Italy

Executive summary

SPICE is an EU H-2020 project dedicated to research on novel methods for citizen curation of cultural heritage through an ecosystem of tools co-designed by an interdisciplinary team of researchers, technologists, and museum curators and engagement experts, and user communities. This technical report D4.5 presents the initial results of Task 4 of Work Package 4: "Provenance and Process analysis layer". The focus of this deliverable is a collection of requirements for provenance and process analysis to be later implemented by the Linked Data infrastructure under development – the SPICE Linked Data Hub. The analysis of this deliverable considers state-of-the-art literature and validates its findings against the needs of museums curators, both from the perspective of the SPICE museums and the pilot studies under development. The objective is to devise the requirements that a Linked Data Infrastructure should support in order to streamline the collection and delivery of provenance metadata, as well as monitoring capabilities necessary for advanced process analytics, for the benefit of data managers and application developers. Alongside discussing the requirements from the user needs perspective (the museum curators and the citizens), we also elaborate on how they could be implemented on the Linked Data Hub. This is an interim report that will be followed by D4.6 in the third year of the project, which will be dedicated to implementation and case studies.

Document History

Version	Release date	Summary of changes	Author(s) -Institution
V0.1	09/02/2022	First draft released	Enrico Daga - OU
V0.2	01/03/2022	Detailed structure of the document, with feedback from WP2 and WP7	Enrico Daga – OU, UH, Aalto, AAU, UNITO
V0.3	15/03/2022	Content developed	Enrico Daga (OU), Paul Mulholland (OU)
V0.4	7/04/2022	Internal review	Belen Diaz Agudo (UCM), Luis Emilio Bruni (AAU)
V0.5	25/04/2022	Incorporated feedback from internal review	Enrico Daga (OU), Paul Mulholland (OU)
V1.0	26/04/2022	Final version submitted to REA	UNIBO

Table of Contents

Project information 3
Project contacts 3
SPICE consortium 3
Executive summary 4
Document History 5
1. Introduction 7
2. Related work 7
3. Citizen Curation: requirements for provenance and process analysis 9
4. Citizen curation: a *data journeys* perspective 12
5. Conclusions 14

1. Introduction

SPICE is an EU H-2020 project dedicated to research on novel methods for citizen curation of cultural heritage through an ecosystem of tools co-designed by an interdisciplinary team of researchers, technologists, museum curators and engagement experts, and user communities. The objective of the Work Package 4 is to research on the application of Linked Data principles to connect cultural objects, collections, and citizen contributions, into an infrastructure for interoperability and knowledge exchange within citizen curation activities. While the WP aims at providing the infrastructure for interoperability within the project, by doing that, its goal is researching on a social media infrastructure that can support museums and technologists with:

(1) privacy-aware content sharing methods, so that museums can expose their catalogue and digital assets in a safe and controlled data environment;

(2) methods for expressing and reasoning over fine-grained policies and constraints associated to digital assets;

(3) linking assets and metadata to support search and discovery capabilities (on top of a secure and controlled data environment); and

(4) content provenance, usage tracing, and monitoring in order to support large scale analyses of user-generated, (anonymised) content.

This technical report D4.5 presents the initial results of Task 4 of Work Package 4: "Provenance and Process analysis layer". The focus of this deliverable is a collection of requirements for provenance and process analysis to be later implemented by the Linked Data infrastructure under development – the SPICE Linked Data Hub. The analysis of this deliverable considers state-of-the-art literature and validates its findings against the needs of museums curators, both from the perspective of the SPICE museums and the pilot studies under development (see deliverables 2.3 and 7.5).

Specifically, we analyse requirements towards supporting representation and reasoning about provenance information, for instance, to raise potential copyright violations or recommend fair practices for the reuse of derived assets. In addition, we explore potential support for monitoring the use made of asset within the distributed platform, supporting museum curators (that we call *custodians*), end users, and application builders in dealing with potential inconsistencies.

The objectives of this deliverable are: (a) to devise the requirements that a Linked Data Infrastructure should support in order to streamline the collection and delivery of provenance metadata, focusing on citizen contributed content; and (b) to define the monitoring capabilities necessary for advanced process analytics, for the benefit of citizen curation applications. This is an interim report that will be followed by D4.6 in the third year of the project, which will be dedicated to concrete, supported use cases.

The next Chapter provides an overview of established research in provenance and process representation and reasoning. Chapter 3 introduces a general use case scenario, devised by abstracting the scenarios of SPICE pilots, and provides an overview of the core requirements for provenance and process analysis. Chapters 4 is focused on discussing requirements for provenance and process analysis from the point of view of *data journeys*, and considering the potential SPICE Linked Data Hub. Chapter 5 summarises our contributions and concludes the deliverable.

2. Related work

As we increasingly rely on complex systems to support the management of cultural heritage collections as well as digitally-mediated systems to enable innovative engagement applications, it becomes important to equip underlying infrastructures with means for monitoring, capturing, and explaining what users do with those systems. Recently, since artificial intelligence applications are built using complex data science

workflows, it has become urgent th~~eat we~~ need to understand, integrate, and explain such workflows comprehensively and at a higher level of abstraction. The ACM Principles for Algorithmic Transparency and Accountability[i] emphasise the notions of awareness, audibility, data provenance, and explanation. These principles reflect the idea that stakeholders should have access to and understand what is going on in complex applications and the data models that are part of them at the appropriate level of abstraction. Fortunately, the Semantic Web community developed a variety of knowledge representation formalisms to capture fundamental elements of data science workflows, such as provenance, data flows, and high-level activities. Unfortunately, data science workflows are very complex and, although models and techniques for representing code as a data graph exist[ii], collecting and reasoning on high-level, compact representations is still an open problem.

We choose to identify a key concept as a driver for our analysis: the broad notion of *data journeys*. This concept is receiving increasing attention in the data studies literature. Specifically, the recent edited volume by Leonelli and Tempini[iii], which brings together different domain perspectives, from plant phenomics to climate data processing, and presents them through the lens of data journeys. As such, data journeys incorporate the two fundamental dimensions of our analysis: provenance (what is the lineage of a given data object) and process (what type of operations/actions are the objects involved with).

Fundamentally, here we argue that the journey a citizen curation object goes through, its *lineage* or *provenance*, is a powerful unit of analysis for making sense of citizen curation applications. Therefore, we survey related work in provenance and process analysis.

The term citizen curation object is used to refer to any digital resource used or created through the citizen curation process. This includes: (i) digital representations of artworks and their metadata and museum labels, (ii) resources that guide the citizen curation activity (e.g. quizzes, interpretation exercises) and (iii) the results of the activity (e.g. citizen answers, stories, interpretations). The data and metadata associated with the results of citizen curation activities (e.g. citizen answers to question plus metadata associated with the author (e.g. their identity and community membership) and content the of the activity (e.g. the text and extracted featured such as its sentiment and values) is what D3.5 refers to as interaction data.

Provenance is a well-established notion in museum curation, where it is related to ensuring the quality and lineage of an object as part of the acquisition management phase. This idea has been borrowed by information science research, that reformulates it as the problem of describing how a certain information object has been produced, who is responsible for it, and associated usage requirements. This notion is well-known in other areas such as digital library research, where the importance of understanding the context in which catalogue metadata is being produced, and the impact that such background has in how the catalogued items are perceived by the reference community, is well-understood[iv]. The need to understand the provenance of data has been well documented in the data management[v] and web[vi] literature, which has investigated approaches for representing, extracting, querying, and analysing provenance information. This includes considering the people as content creators on the web, and advocating for integrating this feature at the core of the semantic web[vii]. Indeed, the importance of understanding provenance for web information led to the W3C Prov standard for provenance interchange[viii] as well as the recent Coalition for Content Provenance and Authenticity[ix]. However, a solution which results in a wide application of provenance information as a means for tracing content use on the web is still to come. We refer to the two surveys cited above for more information about provenance systems in semantic web research. It is in this declination that provenance becomes a relevant concept for SPICE, where the assets produced by citizen curation activities are supposed to be managed as first-class object in museum archives.

The need to tie data to the workflow that generated it has been recognised in the scientific workflow community[x]. An essential contribution of this line of work is that, for better supporting process analytics, different granularity levels of activity representations (e.g. high-level tasks in the domain instead of command-line tool parameters) should be captured[xi,xii]. These representations can then be bundled together with the corresponding data assets and other documentation, creating a *research object*[xiii] that can be published using web standards[xiv]. However, most data science programs are not expressed with such

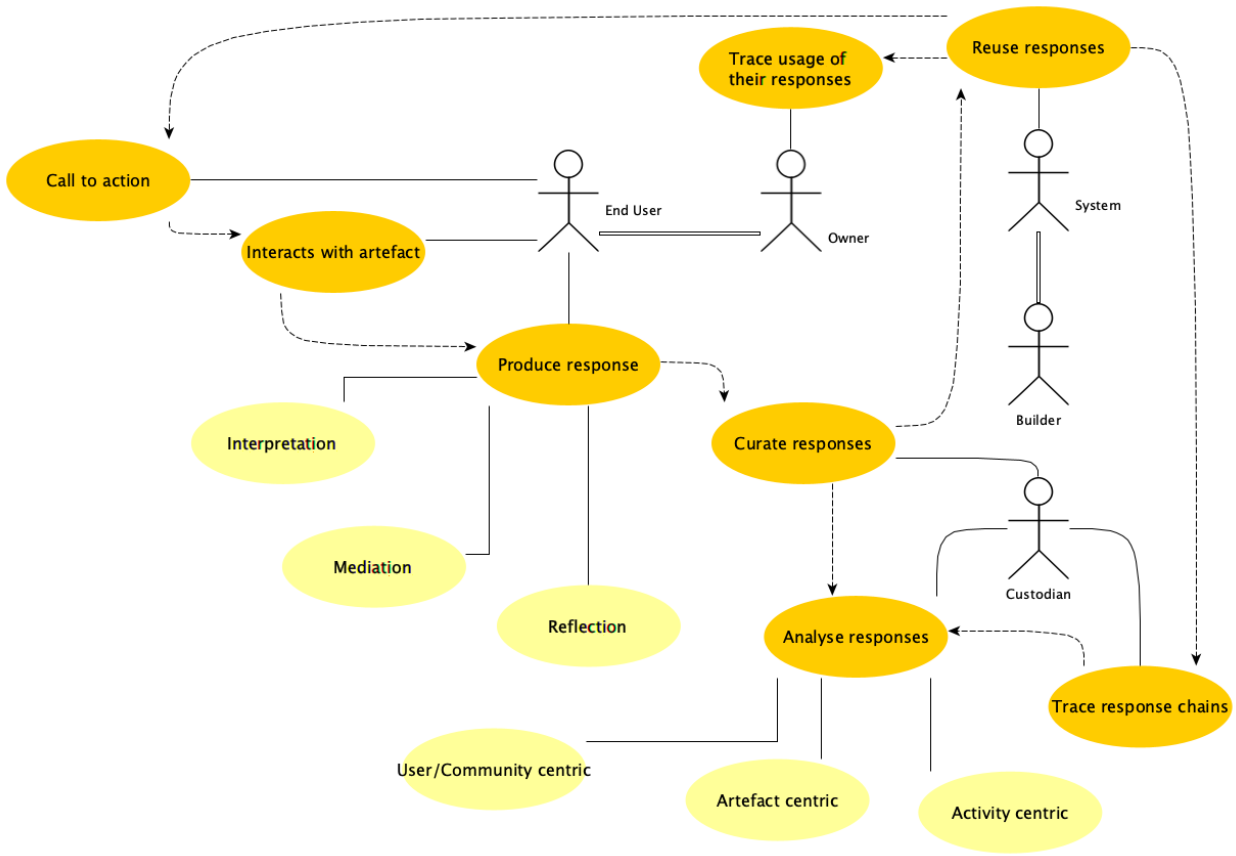
workflow formalisms, and it is unrealistic to assume that those could play a key role in the context of cultural heritage applications.

A complementary perspective relates to capturing *data flows* directly from within the applications. These approaches work for workflow systems where tasks and their dependencies are systematically defined, or for applications that use flexible programming languages. It is the case for data science methods[xv]. Therefore, to perform analytic and assistance, a variety of work has looked at extracting high-level workflow-like representations from code or logging information. Tessera, for example, has looked at extracting high-level tasks from logs of exploratory data analyses[xvi].

In SPICE we are considering citizen curation as a family of methods reusing a catalogue of components via guide-templates (scripts) which encode variants of the interpretation/reflection loop across different museums, engagement activities, and target communities (see deliverables 2.3 and 2.4). In essence, this viewpoint can strongly benefit from a provenance and process analysis perspective, which is new to the domain, and that is the object of this deliverable.

3. Citizen Curation: requirements for provenance and process analysis

The scenario below provides motivation and background for analysing requirements in provenance and process analysis. The scenario and this deliverable particularly focus on citizen contributions rather than provenance in general. The scenario is an abstraction produced by generalising across elements of the five SPICE case studies (see D7.5).



Use the Scenarios as baselines for devising analytics requirements (WP2/WP7)

The Citizen Curation process usually begins with the citizen in the role of end-user receiving some prompt or "call to action" that involves interacting with the collection or exhibition objects of the museum in physical

and/or digital form. This could take place either on-site or off-site. The specific calls to action differ across the case studies. In DMH, this may involve selecting an artefact to interact with in the Pop-up Museum. In GAM, this may involve selecting an artwork with which to tell a story using Emojis. In the case of Hecht, this may involve selecting an artwork, taking photographs of it and expressing an opinion in response to one of more questions. In the case of IMMA, this may involve selecting a script associated with a theme and answering the questions within the script. In the case of MNCN, this may involve selecting a puzzle, solving the puzzle and expressing opinions related to the topic. In all of these cases, the response produced by the citizen can be thought of as a form of interpretation in which they share what they think or know about the exhibit and as well potentially broader topics and themes that the exhibit is used to illustrate. For detailed case specific user journeys see D2.3.

The call to action can also be an invitation to produce an activity to be carried out by other citizens, rather than to take part in a previously created activity. For example, in IMMA this could involve the citizen creating a script to be used by other visitors. In MNCN, this could involve school children creating a new puzzle that could be used on future school or informal visits. This type of response to a call to action can be described as a mediation, as it is intended to mediate someone else's interaction with a museum object or exhibition.

A third call to action can involve evaluating prior interpretations. These could be interpretations made by yourself or by others. This evaluation could lead to a particular type of response such as rating, agreeing, disagreeing or liking other interpretations. This type of activity, where the focus is other's responses rather than the artefact can be described as a form of reflection.

In practice a call to action could comprise one or more of these types of response. For example, a longer, composite call to action could involve creating an activity (mediation), taking part in the activity (interpretation) and later evaluating interpretations produced through the activity by yourself and others (reflection).

The resulting responses (whether reflections, interpretations or mediations) can be curated by the museum custodian. One form of curation is to select responses for reuse as part of the museum's public facing resources. For example, opinions expressed in an activity at Hecht may be reused and presented to other visitors as examples of opinions similar or different to their own. DMH or GAM visitors may access stories authored by previous visitors. IMMA or MNCN may access interpretations and/or mediations created by other visitors. The process of curating responses for reuse may be carried out manually by the custodian, may be system automated by the recommender component, or potentially a combination of both.

The museum custodian can also use the citizen responses for different types of analysis. This can be used to gain a better understanding of the museum's audience and what they think about the museum's current public offering. Analysis can be centred on activities, artefacts or users. Activity-centric analysis investigates how citizens respond to particular activities or types of activity. This could involve, for example, exploring the range of answers given to a question or puzzle, or the emotional content of responses produced in a storytelling activity. Artefact-centric analysis investigates the nature of responses associated with a particular artwork. This could consider the different ways in which the artwork is interpreted. User-centric analysis investigates the range of responses produced by particular users or user groups. This could involve comparing the interpretations of citizens from different community groups or having different demographic profiles (for more information on communities, see D3.5).

Activity, artefact and user centric analysis can also be used in any combination. For example, analysis could consider how responses to an artwork are affected by the nature of the activity, combining activity and artefact-centric analysis. Similarly, analysis could consider how different community groups respond to the same activity and/or artwork.

Analysis can also focus not only on single responses, but interconnected chains of responses. For example, an opinion expressed by one visitor may be reused in other activities, in which another visitor agrees or disagrees with the opinion expressed. Analysis could consider, for example, the volume of responses within a chain and the level of agreement or disagreement they contain, as also suggested in D2.4.

The citizen curator takes the role of owner since they are the authors of any response they have contributed to. As well as maintaining rights over their contribution the owner may wish to trace its reuse. For example, a student may wish to see answers to a puzzle that they have authored, or a visitor expressing an opinion may wish to see how others have reacted to their opinion. Although the citizen curator may retain ownership of their response, the custodian has responsibility for the museum's public offering, which may contain citizen responses. Essentially, the custodian has an editorial responsibility even for content that was not authored by the museum.

Requirements relate to two main actors: the citizen curator as the author and owner of the response, and the custodian who has editorial responsibility over the publication of responses. Some of the requirements below may or may not apply depending on the editorial model adopted, for example, whether the custodian as editor can unilaterally edit a citizen curator's contribution, or whether the citizen curator needs to approve or be informed of the change.

Citizen curator:

- R0 – Intellectual property. Content is contributed under an agreed copyright and terms of use framework (established by the museum and/or the citizen).
- R1 – Understand moderation. Understand the publication status of their responses, e.g. whether the response was held for moderation and the outcome.
- R2 - Monitor use. Monitor use made of their response in activities and public presentations of their citizen response.
- R3 - Monitor response chains. Monitor response chains in which their response features and how other citizens reacted to their response.
- R4 – Monitor editorial changes. Monitor and potentially approve editorial changes made by the custodian to their contributions

Custodian:

- R5 – Publish responses. Convey the publication status of responses to their authors, e.g. whether under review or published
- R6 – Monitor contributions. Monitor the suitability of responses for public presentation
- R7 – Monitor changes. Monitor changes to citizen contributions and potentially citizen approval of editorial changes
- R8 – Analyse activities. Monitor responses to different types of activity
- R9 – Analyse artefacts. Monitor responses to different types of artefacts
- R10 – Analyse communities. Monitor responses by different user communities

The editorial model may change across the various use cases but the common requirement is that actions on content need to be traced and decorated with provenance information, specifically, who is the creator and owner of the produced content and what pre-existing content has been referenced or reused.

To conclude this section, we note how many of the requirements in this list are related to issues that we identified at an early stage of the project (see Deliverable 4.1). Table 1 summarises how these new requirements are related to the ones of D4.1 and, in addition, to the ones expressed in the peer-reviewed article[xvii].

Table 1. Relation between provenance and process analysis requirements and requirements expressed in D4.1 and in the ACM JOCCH article.

R	Description	D4.1 Requirement	Target	JOCCH Requirement	JOCCH description
R1/ R2/ R3/ R4	Understand Moderation; Monitor use; monitor response	[AnalyseUsage]	access and usage of my data	owner:know	

	chains; monitor changes.				
R0	Intellectual property	[ExpressCopyright]	the copyright associated with digital assets in my collection	custodian:copyright	Declare the intellectual property (copyright) associated with the assets
R0	Intellectual property			enduser:copyright	Understand information about copyright and terms and conditions associated to the digital assets in an accessible way
R0	Intellectual property	[GrantRecovery]	terms of use granted	builder:viewterms	Review the terms of use associated to a digital assets
R2/ R3	Monitor use/response chains	[MonitorAccess]	access to my data	owner:know	Ability to know how the digital asset is used
R6/ R7	Monitor contributions/changes			custodian:monitor	Monitor content integrated into the archive to raise issues with relation to copyright infringement or privacy law
R3	Monitor response chains			custodian:usage	Monitor, trace, and analyse the usage of the assets by third parties
R0	Intellectual property	[MultipleRightsAspects]	that multiple subjects hold copyrights on different aspects of the digital asset	custodian:copyright	Declare the intellectual property (copyright) associated with the assets

4. Citizen curation: a *data journeys* perspective

In this section we provide a discussion of the requirements from the perspective of a *data journey*.

The notion of data journey has been discussed in the data studies literature. Specifically, as discussed in Section 2, Leonelli defined it as the ``*movement of data from their production site to many other sites in which they are processed, mobilised and re-purposed.*". The work in data studies emphasises the difficulty of understanding data journeys empirically because of a multitude of perspectives. Our definition has the objective of being consistent with the one of Leonelli but also to relate with the literature from Web semantics, specifically, data provenance. Hence, we introduce a layered semantics perspective to the definition of data journeys[xviii]:

a Data Journey is a multi-layered, semantic representation of a data processing activity, linked to the digital assets involved (code, components, data).

Thus, a journey is multi-layered, as to allow a multiplicity of perspectives that can be overlaid to describe the process. This multiplicity can help to capture (parts of) the context around a data journey while still allowing for computational analysis to be performed. Hierarchical, because any useful representation needs to be linked to the concrete assets involved, either directly or via intermediate abstractions. In this work, we conceptualise data journeys in the following layered structure:

- Resources: resources used in the data journey such as artwork images, metadata records, data sources, licencing information, and terms of use, each one identified by a Linked Data entity URI.
- Event Logs: human readable descriptions of operations produced by runtime processes, for human auditing, such as a Web server requests or entries in a distributed ledger. For example, a citizen curation activity that generates a response.
- Datanode Graph: as defined in[xix], a graph of *data-to-data* relationships, such as variables, imported libraries, and input and output resources. Such abstraction provides a structure of the data flows, abstracting from issues such as control flow, and focusing on *data-to-data* dependencies.
- Activity Graph: a graph of high-level activities, inspired by the notion of Workflow Motifs[xx]. In the context of citizen curation, these can be specialisations of the general scenario introduced in Section 3 (call to action, generates response, reuse response, etc...).

While the first two components pre-exist the data journey, i.e. they do not pertain to the *knowledge level*, the remaining represent two distinct, although interconnected, representation layers. It is worth noting that our definition is open-ended and allows for multiple (even alternative) perspectives to co-exist.

In our reference scenario, citizen responses may be generated by users interacting with a mobile application, when the underlying system (the app itself) generates a new event log referencing the artifact, the type of engagement activity, and the response. After that, a curator could pick up the generated content and modify it, to remove, for example, personally identifiable information. Another citizen may receive a notification, via another citizen engagement system in SPICE, asking to react to that original response. The new user comments with an emoticon, and the underlying infrastructure record the new event.

Event logs deals with the problem of ***capturing provenance*** information. In the SPICE Linked Data Hub, we are developing an activity monitoring layer that has the purpose of recording events from connected citizen curation applications, linking catalogued artifacts with citizen responses, and make them reusable for analysis. The backbone representational layer will be the established W3C Prov-O data model[xxi]. The model can be further extended covering the specificity of citizen curation artifacts and activities. Crucially, the model should cover metadata including ownership and terms of use of the involved assets *at the time* of the event. It is worth noting how such representational layer is agnostic with respect to the underlying technology. Event logs described as such could be stored in a traditional relational database, in a graph database, or in a blockchain (generating non-fungible tokens - NFT).

However, such event logs constitute information that needs to be reused by further activities, from both citizens and museum practitioners. The Linked Data layer should ***provide provenance*** information as part of its service provision, allowing data managers ~~to~~ (curators, developers) to review the information via the Linked Data Hub interface, as well as providing such information to authorized third-party applications (primarily, the one triggering the event). This can be achieved by extending the HTTP protocol and provide an additional, dedicated HTTP response header. Such header may link any served content to the related provenance description. In summary, for each request to the LDH Web API, an equivalent request will be possible to a *provenance* endpoint. Such endpoint will provide provenance information, tailored to the credentials of the requesting system. For example, by requesting a specific citizen response, the citizen curation application will be able to know who produced that response, whether it was authored and by whom. Similarly, when a citizen curation script uses an artifact image, the provenance layer will describe how that metadata record entered the LDH, and link to the original source, being it the museums' Website or the

CMS. Crucially, provenance information will include the usage policies applicable to that context so that applications can adapt and mediate intelligently with their users.

Provenance resources such as data source or terms of use may be associated to a whole collection of linked data entities (such as catalogue metadata). However, catalogue-level terms of use may not be applicable to all items in the same way, as specific artwork images, for example, may have different terms of use. In addition, HTTP requests may vary and refer to multiple linked data entities. One problem that arises relates to the necessity to manage a wide amount of provenance information, potentially for each possible request/response. Duplicating this information for each possible entity URI (or event request) in the Linked Data space would be inefficient (in terms of resources) and overwhelming for the data managers. **Deriving provenance** information requires to develop an intelligent method that is able to propagate relevant data from neighbour objects, when applicable (e.g. considering the terms of use of the catalogue metadata for all its records, without the need of duplicating information). Such system should be able to provide provenance information *on-demand*.

The editorial workflow sketched in Section 3 can only be managed if the infrastructure is capable of recording what citizen curators and custodians do with the managed content. **Capturing usage activities** is a crucial requirement for citizen curation. We discussed *Event Logs*, above, as one foundational layer of data journeys. Usage analytics is a wide area of application on the Web, whose main domain is marketing and advertisement. In our case, the linked data layer needs to record logs about citizen curation activities, covering the whole scenario depicted in Section 3. Such records should be made available to the curators for analytics purposes. Crucially, the Linked Data Hub should be able to provide a high-level representation of how a certain asset (artifact, image, etc...) is being used, by whom, and for what purpose, in SPICE applications.

The resulting workflows will generate a collection of activities associated with assets and related metadata. To support this complexity, systems need to **reason on composite terms of use**. Citizen curation applications may generate composite objects, including images of artefacts, curators' notes (e.g. questions of a slow looking activity), and citizen contributed content. Applications should make users aware of the difference in terms of use associated with each one of them. Potentially, an intelligent system could raise issues in relation to intended use (using an exemplary workflow to verify agreement with current policies). When terms of use affect access control, relevant users should be notified and instructed on what type of actions are needed to ensure a continued availability of resources. For example, when an owner changes the terms of use of an image, they should be notified that there are applications having rights to access that image for a purpose that should not be allowed anymore. In this case, the owner may decide to either revoke the permission or restore the original policies. Similarly, curators shall know if a citizen does not want their content to be used anymore, and such changes should be made available to all the users and citizen curation applications affected.

By representing the events in the linked data layer as data journeys, we can support analytics covering different dimensions, an important need of museum curators in SPICE (see Requirements R8, R9, and R10). Curators can use the data journeys in order to explore the responses to a given artifact, responses of a given community, or how different activities relate. Such representation can be leveraged by an analytics dashboard able to support curators in exploring the contributions from a multiplicity of perspectives.

5. Conclusions

In this deliverable, we analysed citizen curation from the perspective of provenance and process analysis. We designed a general editorial workflow for managing citizen contributions in the context of citizen curation activities, considering the heterogeneity of use cases under development in WP7. From this scenario, we derived a set of user requirements that a Linked Data Infrastructure should support in order to streamline the collection and delivery of contributed content, to benefit of data managers and application developers. In addition, we focused on discussing requirements for provenance and process analysis from the point of view of *data journeys*, and considering the potential SPICE Linked Data Hub. This is an interim report that will

be followed by D4.6 in the third year of the project, which will be dedicated to how they are implemented on the Linked Data Hub, referencing concrete, supported use cases.

---

[i] Council, ACM US Public Policy. "Statement on algorithmic transparency and accountability." Commun. ACM (2017).
[ii] Daga, Enrico, Aldo Gangemi, and Enrico Motta. "Reasoning with data flows and policy propagation rules." Semantic Web 9, no. 2 (2018): 163-183.
Abdelaziz, Ibrahim, Kavitha Srinivas, Julian Dolby, and James P. McCusker. "A Demonstration of CodeBreaker: A Machine Interpretable Knowledge Graph for Code." In ISWC (Demos/Industry). 2020.
[iii] Leonelli, Sabina, and Niccolò Tempini. Data journeys in the sciences. Springer Nature, 2020.
[iv] Chowdhury, Gobinda. "From digital libraries to digital preservation research: the importance of users and context." Journal of documentation (2010).
[v] Herschel, Melanie, Ralf Diestelkämper, and Houssem Ben Lahmar. "A survey on provenance: What for? What form? What from?." The VLDB Journal 26, no. 6 (2017): 881-906.
[vi] Moreau, Luc. The foundations for provenance on the web. Now Publishers Inc, 2010.
[vii] Harth, Andreas, Axel Polleres, and Stefan Decker. "Towards a social provenance model for the web." (2007).
[viii] Groth, Paul, and Luc Moreau. "PROV-overview. An overview of the PROV family of documents." (2013).
[ix] https://c2pa.org
[x] Pérez, Beatriz, Julio Rubio, and Carlos Sáenz-Adán. "A systematic review of provenance systems." Knowledge and Information Systems 57, no. 3 (2018): 495-543.
[xi] Garijo, Daniel, Pinar Alper, Khalid Belhajjame, Oscar Corcho, Yolanda Gil, and Carole Goble. "Common motifs in scientific workflows: An empirical analysis." Future Generation Computer Systems 36 (2014): 338-351.
[xii] Khan, Farah Zaib, Stian Soiland-Reyes, Richard O. Sinnott, Andrew Lonie, Carole Goble, and Michael R. Crusoe. "Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv." GigaScience 8, no. 11 (2019): giz095.
[xiii] Belhajjame, Khalid, Jun Zhao, Daniel Garijo, Matthew Gamble, Kristina Hettne, Raul Palma, Eleni Mina et al. "Using a suite of ontologies for preserving workflow-centric research objects." Journal of Web Semantics 32 (2015): 16-42.
[xiv] Soiland-Reyes, Stian, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo et al. "Packaging research artefacts with RO-Crate." Data Science Preprint (2021): 1-42.
[xv] Subramanian, Krishna, Nur Hamdan, and Jan Borchers. "Casual notebooks and rigid scripts: Understanding data science programming." In 2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp. 1-5. IEEE, 2020.
[xvi] Yan, Jing Nathan, Ziwei Gu, and Jeffrey M. Rzeszotarski. "Tessera: Discretizing Data Analysis Workflows on a Task Level." In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-15. 2021.
[xvii] Daga, Enrico, Luigi Asprino, Rossana Damiano, Marilena Daquino, Belen Diaz Agudo, Aldo Gangemi, Tsvi Kuflik et al. "Integrating citizen experiences in cultural heritage archives: requirements, state of the art, and challenges." ACM Journal on Computing and Cultural Heritage (JOCCH) 15, no. 1 (2022): 1-35.
[xviii] Consistent with Daga, Enrico, and Paul Groth. "Data journeys: knowledge representation and extraction." (Under review)
[xix] Daga, Enrico, Mathieu d'Aquin, Aldo Gangemi, and Enrico Motta. "Propagation of policies in rich data flows." In Proceedings of the 8th International Conference on Knowledge Capture, pp. 1-8. 2015.
[xx] Garijo, Daniel, Pinar Alper, Khalid Belhajjame, Oscar Corcho, Yolanda Gil, and Carole Goble. "Common motifs in scientific workflows: An empirical analysis." Future Generation Computer Systems 36 (2014): 338-351.
[xxi] https://www.w3.org/TR/prov-o/