Applying Effective Data Modelling Approaches for the Creation of a Participatory Archive Platform

JULIEN ANTOINE RAEMY, University of Basel, Switzerland

The development of a participatory archive platform such as the one being carried out for the PIA research project requires a flexible infrastructure allowing genuine data curation and a robust underlying data model. A strong assumption to achieve this is to primarily leverage Linked Open Usable Data (LOUD) standards, such as IIIF, Linked Art or the Web Annotation Data Model, which help in the dissemination and reuse of cultural heritage resources as well as contributing that digital humanities initiatives become more sustainable.

CCS Concepts: • **Theory of computation** → **Data modeling**; • **Information systems** → **Semantic web description languages**; **RESTful web services**; **Ontologies**; **Thesauri**; **Digital libraries and archives**; • **Software and its engineering** → **Interoperability**.

Additional Key Words and Phrases: Citizen Science, Cultural Heritage, Digital Sustainability, International Image Interoperability Framework (IIIF), Linked Art, Linked Open Usable Data (LOUD), Schema.org

1 INTRODUCTION

The digitisation of cultural heritage resources by memory organisations, such as libraries, archives and museums, has greatly benefited the research community, enabling them to answer previously elusive research questions, to raise entirely new ones, and to engage in critical reflection on the consequences of digital transformation.

While cultural heritage resources often have the advantage of being in the public domain or licensed for digital processing and publication, the challenges of accessing - which is ultimately the purpose of preservation - and reusing these outputs in the long term are generally not guaranteed because infrastructures to ensure this type of process are lacking [10]. Although there are many institutional or field-specific repositories, generally only a part of the research data is preserved, usually bound to a Data Management Plan (DMP), if not frozen.

The FAIR[1] Principles [23] can help digital humanities to achieve some measure of digital sustainability and keep these data principles as a common thread. However, this is not enough to achieve the flexibility that an infrastructure and the underlying data model should have with regard to sustainability.

One of the strategies to have a constant maintenance of research data in digital humanities is to leverage the architecture of the Web [9], specifically using Linked Open Data (LOD) approaches and ontologies such as CIDOC-CRM [4, 12].

This paper will look at how the human dimensions around the project titled *Participatory Knowledge Practices in Analogue and Digital Image Archives* (PIA) can impact data modelling as well as identifying the requirements for developing a participatory archive platform where web-based standards play a key role in the dissemination and re-usability of knowledge [20].

2 PARTICIPATORY KNOWLEDGE PRACTICES IN ANALOGUE AND DIGITAL IMAGE ARCHIVES (PIA)

PIA is a four-year interdisciplinary research project - running from February 2021 to January 2025 - funded by the Swiss National Science Foundation (SNSF)[2]. The project is conducted by the

[1] Findable, Accessible, Interoperable, Reusable (FAIR)

[2] SNSF. Participatory Knowledge Practices in Analogue and Digital Image Archives. https://p3.snf.ch/Project-193788

Author's address: Julien Antoine Raemy, julien.raemy@unibas.ch, University of Basel, Digital Humanities Lab, Spalenberg 65, CH-4051, Basel, Switzerland.

Institute for Cultural Anthropology and European Ethnology and the Digital Humanities Lab of the University of Basel as well as the Bern University of the Arts.

The purpose of PIA is to design a visual interface with machine learning-based tools to make it easy to annotate, contextualize, organise, and link both images and their metadata, to deliberately encourage the participatory use of archives. It focuses on three very diverse photographic collections of the Swiss Society for Folklore Studies (SSFS)[3], namely:

- SGV_05: *The Atlas of Swiss Folklore* consisting of 256 maps and 1000 pages of commentary. This collection, which hasn't yet been digitised, was commissioned by the SSFS to do an extensive survey of the Swiss population in the 1930s and 1940s on many issues pertaining, for instance, to everyday life, local laws, superstitions, celebrations or labour;
- SGV_10: *Kreis Family* comprises approximately 20,000 loose photographic objects, mostly kept in photo albums, from a wealthy Basel-based family and spanning from the 1850s to the 1970s. The pictures were taken by studio photographers as well as by family members themselves;
- SGV_12: *Ernst Brunner* is a donation of about 48,000 negatives and 20,000 prints to the SSFS from a professional photographer who lived from 1901 to 1979 and who documented mainly in the 1930s and 1940s a wide range of folkloristic themes.

The fact that part of these collections still has to be digitised alongside the project will be an opportunity to investigate the digital transformation processes, the generation of digital surrogates and their associated metadata as well as how the underlying materiality aspects may or may not be preserved.

More specifically, there are two main goals within PIA: the first having a theoretical focus on the systematic analysis and description of the analogue and digital photo archives and the second an implementation focus related to the design of a participatory image archive.

PIA is as much about traditional crowdsourcing as it is about creating a modern Citizen Science platform that enables new uses and helps to streamline the research process of scholars. For this purpose, a user interface and various application programming interfaces (APIs) will be deployed to accommodate various forms of re-use by third parties [17].

3 IDENTIFIED REQUIREMENTS FOR DEPLOYING A CITIZEN SCIENCE INFRASTRUCTURE

Substantial preliminary work, particularly with regard to data cleaning and reconciliation, needs to be carried out prior to truly deploying an infrastructure that ensures content re-usability, whether by the scientific community or for the wider public [16]. Notably, one of the intentions of PIA is to adopt metadata standards and web-based specifications that are well established within the cultural heritage field [13, 15] in the interests of providing a high level of interoperability. To this end, the following list of requirements influencing the data modelling and the overarching architecture of the infrastructure has been identified:

- Management of existing (non-controlled) keywords and the forthcoming folksonomy;
- Creation of controlled vocabularies using the Simple Knowledge Organization System (SKOS) for all PIA-related collections where dedicated terminology can complement existing LOD taxonomies and thesauri, such as the Getty Vocabularies;
- Metadata correction and enrichment through crowdsourcing;
- Validation of metadata created by Machine Learning methods (notably Object Detection and Visual Text Co-Embedding) through the content-based multimedia retrieval system vitrivr [8];

[3] The photo archive of the SSFS can be consulted online at https://archiv.sgv-sstp.ch/

- Annotation and transcription of digital surrogates and born-digital content that are compatible with the W3C Web Annotation Data Model;
- Reconciliation of entities (e.g. agent, concept, place) on the basis of the existing data model and the user-generated metadata;
- Making digitised collections of the SSFS as well as resources uploaded by end users compliant with the Image and Presentation APIs 3.0 of the International Image Interoperability Framework (IIIF) [1, 2, 21];
- Leveraging the IIIF Change Discovery API 1.0, which is an Activity Streams endpoint, to notify amendments to IIIF resources and make them more easily discoverable in a machine-readable manner [3, 7].

The identification of these requirements provides the framework for establishing an infrastructure that could be in line with the FAIR principles. Furthermore, one of the cornerstones of the project is also the synchronisation of the enriched data between the PIA infrastructure and the Data & Service Center for the Humanities (DaSCH), which is in charge of the long-term preservation of the SSFS (digitised or digital-born) photographs and their associated metadata.

4 THE PIA DATA MODEL

In the context of the database migration from Salsah to the DaSCH Service Platform (DSP) and to anticipate the needs of the PIA project, it was decided to add new classes and properties within the RDF-based SSFS ontology, to rename some of them with the `CamelCase` practice, to delete some not or hardly used properties as well as to identify a common denominator with the PIA data model[4], which is still a work in progress [6]. We believe that the best candidate for this mapping, which will be done using the `rdfs:subClassOf` and `rdfs:subPropertyOf` predicates, is Schema.org, a vocabulary providing a very extensive number of classes and properties as well as having the added benefit of being easily indexed by search engines like Google, Bing or Yandex [22].

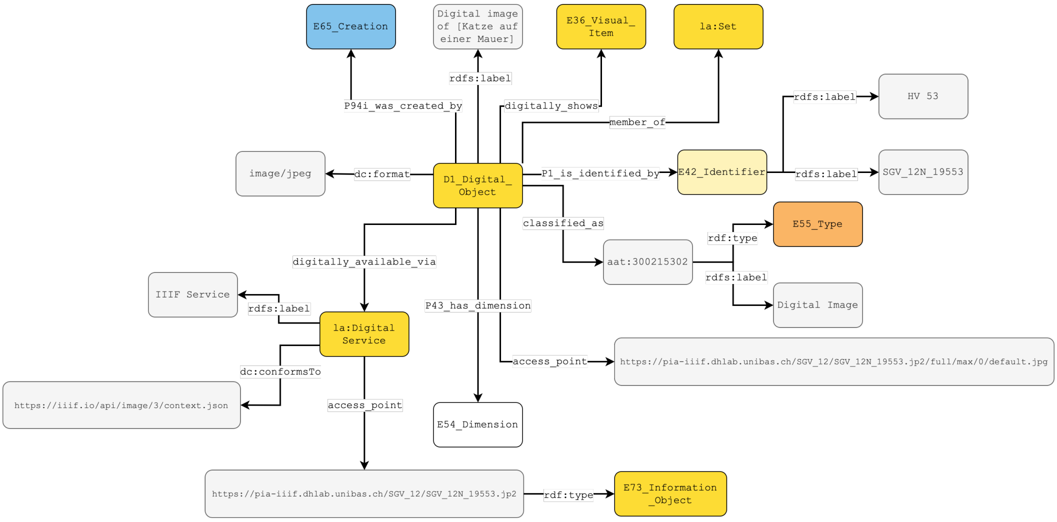

Fig. 1. Partial Representation in Linked Art of *[Katze auf einer Mauer], SGV_12N_19553, https://archiv.sgv-sstp.ch/resource/441788* - object that will have a IIIF Image API 3.0 digital service on the PIA platform

[4] The PIA Data Model GitHub repository is available at: https://github.com/Participatory-Image-Archives/pia-data-model/

If Schema.org is also going to be used within the PIA project and notably by embedding a JSON-LD serialization within the resources' web pages, we also want to provide a representation of the activities surrounding the collections with a focus on their provenance. Thus, Linked Art, which is both a community-based effort and an application profile based on CIDOC-CRM as well as providing an API in JSON-LD [5], will be deployed.

Another advantage of Linked Art is that we can integrate other digital services, such as IIIF (see Figure 1, which is based on the Linked Art Digital Integration component). As announced in the previous chapter, we also want to make the annotations compatible with the Web Annotation Data Model, which can be easily used in conjunction with the IIIF APIs. The three specifications (Linked Art, IIIF, Web Annotatioin Data Model) can be leveraged seamlessly in unison and all enable published data to be easily consumed by both machines and humans. Moreover, these technologies adhere to the design principles of Linked Open Usable Data (LOUD), term coined by Robert Sanderson - which attempts to extend the principles of LOD and who has been working on these specifications [19].

Last but not least, another aspect of the PIA Data Model will be to address specific needs of the SGV_05, SGV_10 or SGV_12 collections in terms of controlled vocabularies. For the SGV_12 Ernst Brunner collection, there is a prominent interest since the photographer had decided to develop his own terminology to structure and classify his photographs. These terms have so far not been included in the metadata. A thesaurus in SKOS has therefore been created and published via SkoHub-Vocabs [14]. It will be refined over the next few months and the items will be connected to this classification for research purposes.

5 CONCLUSION

A flexible data model in digital humanities projects requires seeing the Web as a technology on which other building blocks can be developed (e.g. RESTful APIs) and preferably on LOD for enabling inferences, and even better moving towards specifications that conform to the LOUD design principles for publishing data that is truly usable and which will consequently improve the accessibility and sustainability of digital cultural heritage resources.

The limitation to deploying and maintaining an infrastructure with LOUD standards remains that organisations need to define access persistence policies, notably by assigning cool URIs or persistent identifiers (PIDs) that could at least resolve to alternative web pages containing the associated metadata should they disappear [11, 18].

ACKNOWLEDGMENTS

PIA is made possible thanks to a SNSF Sinergia grant and I would like to thank all my colleagues, especially my supervisors (Peter Fornaro and Walter Leimgruber), the five other fellow PhD candidates (Murielle Cornut, Birgit Huber, Fabienne Lüthi, Florian Spiess and Max Frischknecht) and our software developer (Adrian Demleitner), working on this research project.

Many thanks as well to Yumeng Hou, Doctoral Assistant at the Laboratory for Experimental Museology of the Ecole Polytechnique Fédérale de Lausanne (EPFL), for organising the PhD seminar on Human Factors in Digital Humanities both on site and online on December 2-3, 2021.

REFERENCES

[1] Michael Appleby, Tom Crane, Robert Sanderson, Jon Stroop, and Simeon Warner. 2020. IIIF Image API 3.0. https://iiif.io/api/image/3.0/

[2] Michael Appleby, Tom Crane, Robert Sanderson, Jon Stroop, and Simeon Warner. 2020. IIIF Presentation API 3.0. https://iiif.io/api/presentation/3.0/

[3] Michael Appleby, Tom Crane, Robert Sanderson, Jon Stroop, and Simeon Warner. 2021. IIIF Change Discovery API 1.0.0. https://iiif.io/api/discovery/1.0/

[4] George Bruseker, Nicola Carboni, and Anaïs Guillem. 2017. Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM. In *Heritage and Archaeology in the Digital Age: Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data*, Matthew L. Vincent, Víctor Manuel López-Menchero Bendicho, Marinos Ioannides, and Thomas E. Levy (Eds.). Springer International Publishing, Cham, 93–131. https://doi.org/10.1007/978-3-319-65370-9_6

[5] Emmanuelle Delmas-Glass and Robert Sanderson. 2020. Fostering a community of PHAROS scholars through the adoption of open standards. *Art Libraries Journal* 45, 1 (Jan. 2020), 19–23. https://doi.org/10.1017/alj.2019.32 Publisher: Cambridge University Press.

[6] Adrian Demleitner and Julien Antoine Raemy. 2021. PIA Data Model. https://doi.org/10.5281/zenodo.5142605

[7] Nuno Freire, Enno Meijers, Sjors de Valk, Julien A. Raemy, and Antoine Isaac. 2021. Metadata Aggregation via Linked Data: Results of the Europeana Common Culture Project. In *Metadata and Semantic Research (Communications in Computer and Information Science)*, Emmanouel Garoufallou and María-Antonia Ovalle-Perandones (Eds.). Springer International Publishing, Cham, 383–394. https://doi.org/10.1007/978-3-030-71903-6_35

[8] Ralph Gasser, Luca Rossetto, and Heiko Schuldt. 2019. Multimodal Multimedia Retrieval with vitrivr. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval (ICMR '19)*. Association for Computing Machinery, New York, NY, USA, 391–394. https://doi.org/10.1145/3323873.3326921

[9] Ian Jacobs and Norman Walsh. 2004. Architecture of the World Wide Web, Volume One. https://www.w3.org/TR/webarch/

[10] Andŕe Kilchenmann, Flavie Laurens, and Lukas Rosenthaler. 2019. Digitizing, archiving... and then? Ideas about the usability of a digital archive. *Archiving Conference* 2019, 1 (May 2019), 146–150. https://doi.org/10.2352/issn.2168-3204.2019.1.0.34

[11] Lukas Koster. 2020. Persistent identifiers for heritage objects. *The Code4Lib Journal* 47 (Feb. 2020). https://journal.code4lib.org/articles/14978

[12] Florian Kräutli, Esther Chen, and Matteo Valleriani. 2021. Linked data strategies for conserving digital research outputs: The shelf life of digital humanities. In *Information and Knowledge Organisation in Digital Humanities*. Routledge, London. https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781003131816-10/

[13] Marian Clemens Manz, Julien Antoine Raemy, Béatrice Gauvain, and Vera Chiquet. 2021. Let's talk about standards – A write-up of a discussion on metadata standardization in the Digital Humanities. https://dh-ch.ch/blog/posts/let's-talk-about-standards-a-write-up-of-a-discussion-on-metadata-standardization-in-the-digital-humanities.html

[14] Adrian Pohl and Adrian Ostrowski. 2019. Presenting the SkoHub Vocabs Prototype. https://blog.skohub.io/2019-09-27-skohub-vocabs/

[15] Julien Antoine Raemy. 2020. *Enabling better aggregation and discovery of cultural heritage content for Europeana and its partner institutions*. Master's thesis. HES-SO University of Applied Sciences and Arts, Haute école de gestion de Genève, Geneva, Switzerland. https://doc.rero.ch/record/329698

[16] Julien Antoine Raemy. 2021. Data modelling and Citizen Science: impact of user-generated content within the PIA research project. https://doi.org/10.5281/zenodo.5763545

[17] Julien Antoine Raemy. 2021. The International Image Interoperability Framework (IIIF) APIs as the backbone of scientific and participatory research. https://doi.org/10.5281/zenodo.4890821

[18] Julien A. Raemy and René Schneider. 2019. *Suggested measures for deploying IIIF in Swiss cultural heritage institutions*. White paper. HES-SO University of Applied Sciences and Arts, Haute école de gestion de Genève, Geneva, Switzerland. https://doi.org/10.5281/zenodo.2640416

[19] Robert Sanderson. 2020. The Importance of being LOUD. https://www.slideshare.net/azaroth42/the-importance-of-being-loud

[20] Tobias Schweizer, Lukas Rosenthaler, and Peter Fornaro. 2017. Content-based Interoperability: Beyond Technical Specifications of Interfaces. *Archiving Conference* 2017, 1 (May 2017), 34–38. https://doi.org/10.2352/issn.2168-3204.2017.1.0.34

[21] Stuart Snydman, Robert Sanderson, and Tom Cramer. 2015. The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. In *Archiving Conference*, Vol. 2015. IS&T, Los Angeles, CA, 16–21. https://purl.stanford.edu/df650pk4327

[22] Richard Wallis, Antoine Isaac, Valentine Charles, and Hugo Manguinhas. 2017. Recommendations for the application of Schema.org to aggregated Cultural Heritage metadata to increase relevance and visibility to search engines: the case of Europeana. *The Code4Lib Journal* 36 (April 2017). https://journal.code4lib.org/articles/12330

[23] Mark D. Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18