https://doi.org/10.2352/issn.2168-3204.2019.1.0.2

Archiving AV materials FAIR: An Oral History collection in the repository DANS-EASY

Eliane Fankhauser; DANS; The Hague, The Netherlands

Abstract

It is crucial for both the research community and the public that AV materials are archived in a FAIR manner. This paper is about AV materials contained in DANS's EASY repository, and their depositing and archiving according to the FAIR principles. Using an Oral History collection in EASY as use case,, it is explored to what extent the FAIR principles are followed and where their implementation is challenging. Moreover, tools currently developed to help assess the FAIRness of datasets before or after deposition are introduced.

Audio and video recordings are taking up ever more important roles in our daily lives. It is thus relevant to think carefully about the way to archive these materials sustainably. Findability and accessibility of digital audiovisual (AV) materials are key to keep memory alive and to introduce and present the collections we have to the audience. Both findability and accessibility influence the extent to which materials can be reused. On the other hand, the technical and ethical aspects of archiving AV file formats need to be discussed in detail and supported by common standards which inform the interoperability of the digital preservation in the long term. All these aspects together are part of the so-called FAIR Principles , which stands for Findable, Accessible, Interoperable and Reusable. Nowadays, the FAIR Principles are seen as the standard concept for preserving research resources sustainably.

This paper investigates the FAIR Principles and the aspects which are most relevant for AV materials. A collection deposited in 2016 at DANS's Electronic Archiving System (EASY), a certified Trusted Digital Repository, will be used as use case with which to demonstrate what archiving and publishing in a FAIR manner means in the case of AV materials. Moreover, two online tools will be introduced which can be used to support the process of archiving AV materials FAIRly: the FAIRdat tool and the checklist to evaluate FAIRness of data(sets).

The EASY Repository

The online repository EASY is a service of Data Archiving and Networked Services (DANS), the institute for permanent digital preservation. DANS was founded in 2005 and is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO). Predecessor institutes already concentrated on the preservation of digital research data which adds up to nearly fifty years of experience in the field of digital preservation. EASY went live on the first of January 2007, being a fusion of three systems, the Steinmetzarchief (Steinmetz archive), the e-depot voor de Nederlandse Archeologie (e-depot for Dutch Archaeology) and the Nederlands Historisch Data Archief (Dutch Historical Data Archive) [1]. Currently, the repository holds over 84,000 datasets, a number which increases steadily. In the very beginning of EASY, a majority of the datasets deposited were from the Humanities and Social Sciences. Nowadays, however, researchers from a broad range of disciplines deposit their data in EASY.

The total number of datasets in EASY currently lies at about 84,000, 2,909 of which contain audio visual materials. These datasets are collected under the umbrella term Oral History. It appears that a majority of the Oral History datasets have restricted access rights. Only a small number, namely 294 datasets, are published open access. The reason for this lies in the sensitivity of the material provided – often Oral History datasets contain interviews about WWII – on the one hand, and in the absence of informed consents on the other. Especially interviews recorded in the 1970s to the 1990s, which now are being converted from video cassettes to digital formats, often lack consent and therefore cannot be made available to the public.

AV datasets in EASY typically consist of an overview or jump-off page where general information such as the persistent identifier of the dataset, the citation and a summary of the dataset or the project to which the dataset belongs is displayed. Separate tabs show the metadata ("Description"), the data files ("Data files") and the AV streaming ("Audio / Video"). EASY also makes use of so-called thematic collections (Dutch: *Thematische Collectie*). Collections display an overview of several datasets related to each other. For example, a collection can be established for a project consisting of several interviews. Every interview, together with its corresponding metadata and additional files, is collected in one dataset. All datasets together constitute a collection.

The FAIR Principles

FAIR and its corresponding principles came into existence in 2016 when a broad range of stakeholders in the field of research data management and stewardship came together to discuss the improvement of the reusability of research data. An article about these FAIR Principles, a set of measurable guidelines on FAIRness of datasets [2], stood at the beginning of an ongoing endeavour to improve the discovery and reuse of digital resources and to develop their interoperability within a bigger ecosystem. The FAIR Principles consist of a total number of fifteen guiding principles, every single one of them related to either findability, accessibility, interoperability, or reusability (see Table 1).

Table 1. The FAIR Principles

Findable		
F1		(meta)data are assigned a globally unique and persistent identifier
F2		data are described with rich metadata (defined by R1 below)

F3		metadata clearly and explicitly include the identifier of the data it describes
F4		(meta)data are registered or indexed in a searchable resource
Accessible		
A1		(meta)data are retrievable by their identifier using a standardized communications protocol
	A1.1	the protocol is open, free, and universally implementable
	A1.2	the protocol allows for an authentication and authorization procedure, where necessary
A2		metadata are accessible, even when the data are no longer available
Interoperable		
I1		(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2		(meta)data use vocabularies that follow FAIR principles
I3		(meta)data include qualified references to other (meta)data
Reusable		
R1		meta(data) are richly described with a plurality of accurate and relevant attributes
	R1.1	(meta)data are released with a clear and accessible data usage license
	R1.2	(meta)data are associated with detailed provenance
	R1.3	(meta)data meet domain-relevant community standards

Three years after the first publication on the FAIR Principles, they have become an indispensable part in the discourse of the scientific community, being adopted by a wide range of stakeholders including infrastructure providers, researchers, and disciplinary initiatives. Initiatives like the PARTHENOS "guidelines to FAIRify data management and make data reusable" [3] help the community to implement FAIR data standards and tailor them to the needs of the respective disciplines. Whilst in 2016 the focus lay on the FAIRification of datasets, FAIRness today also applies to software, repositories and other digital resources. However, experience from the past three years has shown that the fifteen current principles have not been formulated neatly enough to accurately measure all aspects of FAIRness. Therefore, a small group of FAIR experts came up with an additional set of metrics [4] to further refine the measurement of FAIRness. On the European level, the Expert Group on Turning FAIR Data into Reality [5] was set up to provide recommendations to facilitate and operationalise FAIR as the standard way of sharing digital resources among the scientific community by 2020. The group published its Final Report and Action Plan [6] in November 2018. Most recently, frameworks for FAIR data and services are addressed by a number of projects funded by the European Commission, among which the FAIRsFAIR project [7].

The FAIR Principles make statements about several levels of data preservation, ranging from the repository in which data is preserved to the machine readability of digital resources and their corresponding metadata. In the context of this paper, the focus will be on those FAIR principles which need special attention when archiving AV materials. These first and foremost concern FAIRness on the level of datasets and data files. It is, however, clear that making AV materials FAIR means that the context in which datasets are stored, i.e. the FAIRness of repositories, also needs to be taken into consideration.

The FAIR Principles in Practice: A Use Case

Erik de Jager is a Dutch documentary filmmaker who says about himself that he has a strong interest in aspects of "everyday life". In 2016, De Jager deposited a data collection titled "Reis van de razzia" (Journey of the Raid) in EASY [8]. The project's aim was to shed light on the cordon that the German army built in Rotterdam and Schiedam in the evening of 9 November 1944. Within two days, more than 52,000 inhabitants of Rotterdam and Schiedam between the ages of 17 and 40 were arrested and transported to Germany to perform forced labor. The beginning of de Jager's description of the collection reads as follows:

The project "The Journey of the Raid" is based on filmed testimonials from men who have experienced the raid and the subsequent journey, to fill a gap in the historiography and to provide insight into the events on the theme "Scope of action of an individual in a society under pressure" [9].

The thematic collection "Journey of the Raid" in EASY consists of a total number of 76 interviews the majority of which are accessible openly. Every interview is deposited as an own dataset, containing (i) the interview to be streamed on the website, (ii) the transcription of the interview in pdf format and (iii) additional information about the route of the journey and earlier records of the men's stories (see Figure 1). Invisible to the regular EASY user, moreover, the datasets also contain informed consents signed by both the interviewee and the interviewer(s). The totality of interviews is added to the thematic collection together with the description of the project on the jump-off page, the metadata page, and more background information about the project. In the following, a selected number of FAIR principles will be described and evaluated against the collection "Journey of the Raid".

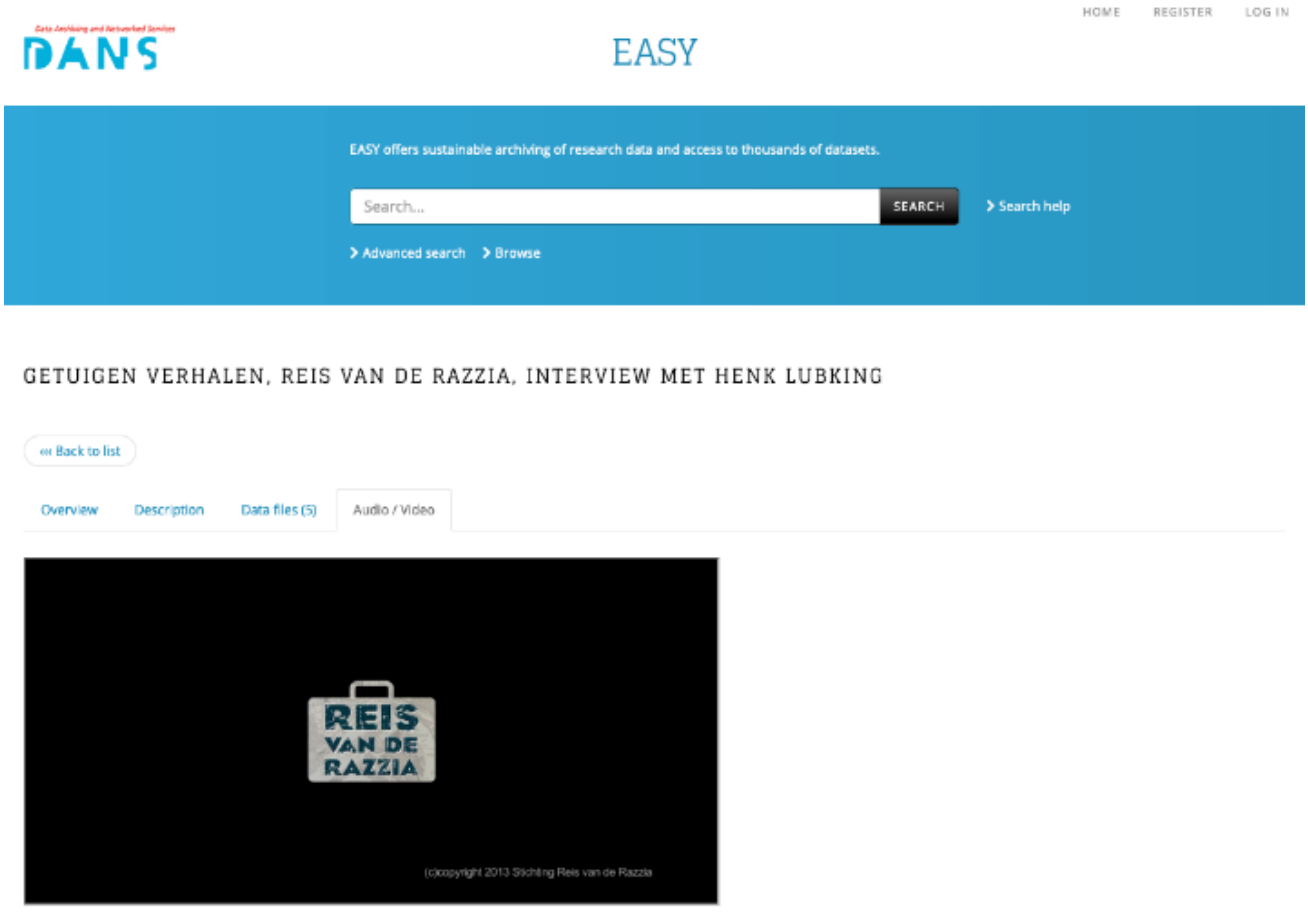

Figure 1. *Streaming a video file in EASY.*

Findable and Reusable

The second principle under "Findable" (F2) concerns the completeness of the metadata: data are described with rich metadata. This principle relates to R1 (meta)data are richly described with a plurality of accurate and relevant attributes. Combined, these principles together focus on whether the metadata and data are findable by humans and machines and whether they are useful. De Jager in his data collection gives a broad range of metadata for all of his datasets, ranging from the title with clear attributions -- it mentions the subject of the collection, the kind of recording and the name of the interviewee--, the creator and the contributors of the datasets, descriptions, and temporal and spatial coverages, to a list of keywords.

Interoperable

The list of keywords in the metadata relates to yet another FAIR Principle: I2 (meta)data use vocabularies that follow FAIR principles. I2 is especially relevant for the usage of so-called controlled vocabularies in the keyword description of datasets. There are hundreds of controlled vocabularies for different disciplines and fields. The Art and Architecture Thesaurus (AAT) [11] is an example of a controlled vocabulary for art history and architecture. In the field of history, to which "Journey of the Raid" belongs, the existence of thesauri is still limited. De Jager provides well-known terms such as "Oral History" next to very specific keywords in Dutch like onderduiken (to abscond). Here the data collection in EASY thus does not comply with the FAIR principle I2, one reason for this being the limited existence of controlled vocabularies in history.

R1.3 (meta)data meet domain-relevant community standards addresses the usage of standards in the respective disciplines in broader terms. Here again, it very much depends on the discipline whether or not community standards and best practices exist and are followed by researchers. Clearly, De Jager's datasets do not follow standards per se because such standards are not yet defined for (Dutch) history of the twentieth century.

One more aspect that is of importance to the archiving of AV materials are the file formats. For the long-term preservation, it is crucial to use sustainable formats. EASY has a list of preferred formats, including an explanation of why the formats listed are chosen to be preferred [12]. De Jager's interviews are .mov files (QuickTime file format). A glance at the list of formats reveals that QuickTime formats are non-preferred. Non-preferred formats are formats that are widely used and which will be moderately to reasonably usable, accessible and robust in the long term. At the time De Jager's interview collection was archived and published, however, the QuickTime file format was still classified as a preferred format in EASY.

Usage licenses and the GDPR

Erik de Jager's collection is one of the few AV collections in EASY that is published open access. There are two preconditions which need to be fulfilled in order to publish datasets open access. First, researchers have to be willing to make their data openly accessible. Second, informed consents tailored to the specificities of audio or video recordings according to the General Data Protection Relation (GDPR) have to be signed by both the interviewees and the interviewers. Ensuring that data and metadata are released with a clear usage

license like the FAIR principle R1.1 (Meta)data are released with a clear and accessible data usage license states is the responsibility of the repository which stores the data (see "More FAIR Principles" below). How open data containing personal information is, however, depends on the statements made in informed consents. For De Jager's interviews such informed consents are signed which clearly state that the interviewee, but also the interviewers, cede all rights. "Journey of the Raid" was published long before the GDPR came into force in May 2018. Currently, the usage licenses in EASY are under review, as they have to be compliant with the GDPR. It is possible that, after this review, "Journey of the Raid" will have a different usage license.

More FAIR Principles

In the above sections, five out of fifteen FAIR Principles were discussed. As mentioned above, it was chosen to focus on those principles which are of special importance for the archiving of AV materials. The remaining ten FAIR Principles are about machine readability and conditions for data archiving which cannot be taken care of by the researchers themselves. As a matter of fact, the majority of these principles, and especially the ones summarized under Interoperability, are covered by Trusted Digital Repositories. Such Trusted Digital Repositories often are certified with the CoreTrustSeal (CTS) which offers a core-level certification. The CTS is granted to those repositories which fulfill the sixteen DSA-WDS Core Trustworthy Data Repositories Requirements [13]. To a considerable extent, these requirements correspond to the FAIR Principles. However, in their current state, they are not fully FAIR compliant. The level of FAIR maturity in repository certification will be evaluated in the upcoming years. EASY, the repository in which "Journey of the Raid" is archived, therefore already shows a certain level of FAIRness.

FAIR Data Assessment tools

Archiving data in a FAIR manner can be a challenging endeavour, especially for researchers who are not data management specialists. This was recognized and reacted to with the establishment of tools to support the assessment of the FAIRness of datasets. Up until the present day, several different tools were developed and tested, most of them today still being beta versions. The working group "FAIR Data Maturity Model" of the Research Data Alliance (RDA) [14] is currently assessing the maturity of the FAIR data principles and also analysing the landscape of FAIR assessment tools. The two tools which will be introduced in this section are developed to help researchers archive and publish their data in a FAIR manner: (i) the FAIR data assessment tool named "FAIRdat" developed by Emily Thomas, Peter Doorn and Eleftheria Tsoupra and (ii) the FAIR checklist for researchers by this author.

FAIRdat [15] is a tool to evaluate the FAIRness of already-deposited datasets. The prototype was established and first published in the summer of 2017 [16]. The tool gives a 5-star rating (see Figure 2) of the Findability, Accessibility, Interoperability and Reusability of a dataset as well as a score of its overall FAIRness. It consists of a number of questions about the dataset which are formulated in accordance with the fifteen FAIR principles. As a support to understand and follow the FAIR principles, FAIRdat can also be used prior to archiving datasets. FAIRdat is still a prototype, awaiting its revision and publishing as a full version.

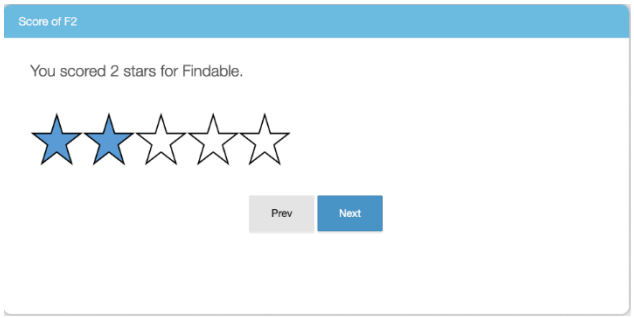

Figure 2. *5-star rating of F2 in the FAIRdat tool.*

The second tool, a checklist for researchers [17], helps to understand the FAIR principles and apply them to data and datasets prior to archiving. It is thought to give an idea about what FAIR data sharing means. The questions are deliberately kept simple. Corresponding paragraphs provided below the questions contain explanations of terms and concepts in a comprehensible language. The assessment covers four levels: (i) the repository in which the dataset will be archived, (ii) the metadata describing the dataset, (iii) the dataset and (iv) the data files. There are thirteen yes-no questions, eleven of which are related to the FAIR Principles. Two questions are about the repository and about open access. After having answered all questions, the user sees an overview with the score. For all questions which were answered with no and for some questions answered with yes, additional suggestions are provided on how to make the dataset more FAIR. The checklist was first introduced to the public in autumn 2018. Just like the FAIRdat tool, it still is a beta version. Both tools described above are easy to use and can assist researchers in raising the level of FAIRness of AV materials.

Conclusion

The evaluation of the FAIR Principles where AV materials are concerned has brought to the fore four aspects which are of great importance. These are (i) providing rich metadata describing the AV materials and additional documentation, like transcriptions, (ii) having signed informed consents of interviewees and interviewers, (iii) making use of community-specific vocabularies and (iv) storing AV materials in sustainable file formats. Erik de Jager's data collection, which in this paper served as a use case for the evaluation of archived data containing AV materials, largely meets the requirements defined by the FAIR Principles in terms of findability and reusability. Aspects of interoperability, namely the use of standardized vocabulary and sustainable file formats, however, are not fully FAIR compliant. It is thus clear that "Journey of the Raid" shows a good overall level of FAIRness, also due to being archived in a Trusted Digital Repository, but that more effort is needed to publish datasets containing AV materials more in line with the FAIR Principles. The two self-assessment tools introduced in this paper can help raise the awareness of FAIR aspects discussed above and improve the overall FAIRness of data.

References

[1] Borgman, Ch. L., Scharnhorst, A., Berg, H. Van den, "Who Uses the Digital Data Archive? An Exploratory Study of DANS." Proceedings of the Association for Information Science and Technology 52, 1, pg. 2 (2015).

[2] Wilkinson, M. D., et al. “The FAIR Guiding Principles for scientific data management and stewardship.” Scientific Data 3, (2016).
[3] PARTHENOS Guidelines to FAIRify data management and make data reusable. http://www.parthenos-project.eu/portal/policies_guidelines.
[4] Wilkinson, M. D., et al. “A design framework and exemplar metrics for FAIRness.” Scientific Data 5, (2018).
[5] Horizon 2020 Commission expert group on Turning FAIR data into reality (E03464). http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3464.
[6] Collins, S., et al (2018). Turning FAIR into reality. Retrieved from https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283 doi: 10.2777/1524
[7] FAIRsFAIR: Fostering FAIR Data Practices in Europe. https://www.fairsfair.eu/.
[8] Jager, E. J. de (2014, September 30): Thematische collectie: Erfgoed van de Oorlog, Getuigen Verhalen, Project 'Reis van de Razzia'. DANS. https://doi.org/10.17026/dans-2a5-ec82.
[9] Ibid.
[10] Stichting Reis van de Razzia. https://stichtingreisvanderazzia.nl/?cat=2.
[11] Art & Architecture Thesaurus Online. http://www.getty.edu/research/tools/vocabularies/aat/.
[12] File formats. https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing/file-formats.
[13] CoreTrustSeal Data Repositories Requirements. https://www.coretrustseal.org/why-certification/requirements/.
[14] FAIR Maturity Model WG. https://www.rd-alliance.org/groups/fair-data-maturity-model-wg.
[15] FAIR data assessment tool. https://www.surveymonkey.com/r/fairdat.
[16] Thomas, E. (2017, June 26). FAIR data assessment tool. Blog post. Retrieved *from* http://blog.ukdataservice.ac.uk/fair-data-assessment-tool/.
[17] FAIR enough? Checklist to evaluate FAIRness of data(sets). https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9IOBoj5GgWqik8KnhtH3B819Ch6lD5KuAz7yn0I0Opw/viewform.

Author Biography

Eliane Fankhauser received her MA in musicology from Utrecht University (2013) and her PhD about polyphonic music in the Netherlands in the late Middle Ages from Utrecht University (2018). At KNAW-DANS she is a project manager in the role of which she currently works in the FAIRsFAIR project. In 2018, she was the coordinator of the Oral History collection in the EASY repository. Together with a small team, moreover, she is working on a FAIR checklist for researchers to measure and evaluate the FAIRness of data.