Enriching Cultural Heritage Data for Research – the Quest for Interoperability in Audiovisual Archives

Mari Wigham
Netherlands Institute for Sound and Vision
The Netherlands

Willem Melder
Netherlands Institute for Sound and Vision
The Netherlands

Roeland Ordelman
University of Twente / Netherlands Institute for Sound and Vision
The Netherlands

ABSTRACT

Digital Heritage and Cultural data are an important information source for researchers in the social sciences and humanities. Research infrastructure projects, such as CLARIAH[1], invest in making such data sets, which are typically maintained by a number of institutions in the Cultural Heritage sector, available to scholars. We see that activities in the fields of digital research infrastructures, and digital infrastructures in cultural heritage are converging, in particular regarding FAIR [2] use of data. This is particularly the case for Linked (Open) Data, which emphasizes the interoperability of data sets - how easily they can be exchanged, combined and used.

Researchers often need to combine and aggregate data from different data sets, sometimes hosted by different institutions. Linked Data principles[2], help to uniquely identify concepts –from basic concepts such as persons, organisations and locations, to complex ones such as works of art and television productions – across different data sets, and describe their properties and the relationships between them in a way that is understandable for both humans and computers. This can assist researchers in finding relevant data and combining them ((semi-)automatically), providing new ways to acquire knowledge and insights.

Cultural heritage linked data in practice. Putting Linked Data principles into practice in cultural heritage collections is, however, far from trivial. Adhering to standards is an important step in making data interoperable. At the same time, to make linked data truly applicable, it is necessary to tackle user scenarios that do justice to the real-life situation in archives that often deal with vast amounts of data, frequently heterogeneous in terms of their medium (text, images, audiovisual), metadata, and accessibility (open, sensitive, copyright).

The CLARIAH Media Suite[3] aims to bring theory and practice together in a digital virtual research environment (VRE) that makes data from media archives (e.g., newspaper collections, radio & television broadcasts, oral history) findable, accessible, interoperable and reusable, and provides tools for researchers to interact with this data (searching, inspecting, analyzing, visualizing, annotating). Interoperability concerns both the data within the Media Suite infrastructure (e.g., linking newspapers to broadcast news), and external data (e.g., enriching broadcast news metadata with information from Wikidata [4]). The Media Suite is an ideal testbed for investigating linked data in a real-life environment.

The heterogeneity of cultural heritage data can be illustrated using the Sound and Vision archive as an example, a Media Suite collection that contains more than 2 million items important to Dutch cultural heritage. The heterogeneity of this archive is because (i) the archive items vary greatly, from wax rolls to costume props to TV programmes; and for each type, different metadata is applicable; (ii) the descriptive metadata has been produced at different times, by different institutions, under different policies and by a range of sources, from humans to AI algorithms; and (iii) the purposes for which the metadata was originally collected are equally diverse - from production information for broadcasters, to technical specifications for archive owners, to historical context for the Sound and Vision museum.

This heterogeneity presents us with a significant challenge - how do we preserve this richness while promoting interoperability? Using standardised schemas such as Dublin Core[5] or schema.org[6] drastically improves interoperability, but at the same time, inevitably leads to loss of data and meaning, as such schemas only cover a portion of the metadata. Some institutions, such as musoW[7] and the Dutch Royal Library[8], address this by using multiple standardised schemas. While this helps the coverage of the data, the schemas won't necessarily have sufficient granularity to cover the detail of the data. Moreover, using multiple models requires that researchers working directly with the linked data be familiar with all these models in order to understand the data. After conversion to linked data, the metadata must be published in such a way that cultural heritage users can find data collections (e.g., in data set registries such as that of the NDE[9], or Europeana[10]), get an overview of what is available in the collection, query or import the data and rely on access to that data continuing to exist into the future. At the same time, the rights of data owners must be protected.

Applying linked data to cultural heritage use cases. Once cultural heritage data has been successfully published as linked data, this does not automatically mean that researchers will be able to utilise it successfully. To work with linked data, a researcher usually needs programming skills and knowledge of Linked Data technologies such as the RDF format[11] and the SPARQL query language[12]. This knowledge is not widespread in the cultural heritage community.

[1] https://clariah.nl
[2] https://www.w3.org/DesignIssues/LinkedData.html
[3] https://mediasuite.clariah.nl/documentation/faq/what-is-it
[4] https://www.wikidata.org/wiki/Wikidata:Main_Page
[5] https://www.dublincore.org/specifications/dublin-core/
[6] https://schema.org/
[7] https://musow.kmi.open.ac.uk/
[8] https://www.kb.nl/bronnen-zoekwijzers/dataservices-en-apis/linked-data-van-de-kb
[9] https://datasetregister.netwerkdigitaalerfgoed.nl/
[10] https://www.europeana.eu/nl
[11] https://www.w3.org/RDF/
[12] https://www.w3.org/TR/rdf-sparql-query/

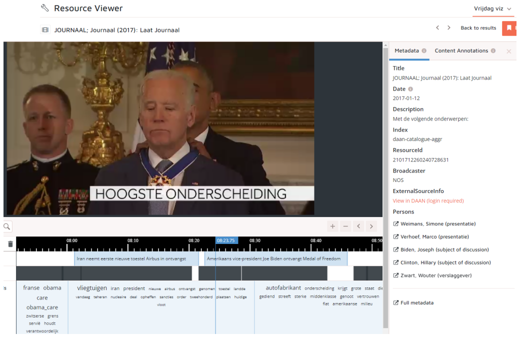
While attention is being paid to growing such skills[13] [1], we assert that linked data should be usable for simpler tasks without requiring such expertise. Advanced users can work with the data directly, while other users can be given tools that find, select and combine the linked data for a given use case, presenting it in a user-friendly way. While concealing the technical details of linked data can aid usability, for research it is essential that data is transparent. The key question posed in our research is therefore: how can we apply linked data in tools that support researchers in finding, understanding and combining data, in a transparent way?

Sustainable Linked Data as a Use Case. In this paper, we first describe the conversion of cultural heritage data to linked data, using the Sound and Vision archive as a specific case. This linked data can be used by advanced users, and also as a source for the additional tools described later. Converting archive metadata to linked data required a concerted effort from a multidisciplinary team, a data modeller to model the data, a copyright specialist to check legal issues, a data architect to design and build prototype infrastructure to retrieve, transform and publish data, and a software architect to implement this infrastructure in a robust and production-ready environment.

We discuss the available metadata schemas, compare these to the Sound and Vision metadata, and justify our choice to develop our own metadata schema to preserve the richness and diversity of the original data. We describe how we handle the heterogeneity problem by delivering different views on our linked data for different use cases. Each view is tailored to a use case, and is therefore less heterogeneous than the entire archive. It is thus more likely that an appropriate standard schema exists to model the relevant data, optimising interoperability for that case. If a use case demands a specific schema (e.g., when delivering data to aggregators), then we can accommodate that. For niche use cases where no suitable schema exists, we can use our own schema. Our infrastructure combines on-the-fly data mapping with profile negotiation[14] to create views. This means that we can support multiple schemas - and use cases - with the same data using the same infrastructure and, within a use case, allow advanced users to choose the most appropriate schema. Example queries are supplied to help the advanced user quickly get to grips with the data.

For archives that contain copyrighted material, filtering of sensitive information and eventual delivery of data under the applicable license is essential. We explain how we ensured this when publishing the data. We outline ongoing work to make the linked data findable, e.g., by using the data set registry of the NDE digital cultural heritage network.

Second, we describe how we exploit Linked Data in a research environment for cultural heritage researchers without the technical skills to "program with data". We present the use case of researchers interested in persons, including the problems such researchers currently face in finding relevant search results for persons in search tools such as the Media Suite. We combine our archive metadata with our GTAA thesaurus[15] and person data in Wikidata to provide richer information about the persons in the archive (see also Figure 1). Further, we explain how the Media Suite uses this information to enhance its search functionality. Autocompletion, disambiguation and query expansion with alternative names provide more –and more relevant– search results, while the enriched information provides more details about the persons appearing in search results, to help the researcher better understand who these persons are and what their role is in the search result (e.g., a guest in a programme, or the producer). The researcher benefits from linked data without bothering with technical details, while transparency is ensured by links to the original source data.


Resource Viewer
Vrijdag viz
Back to results
HOOGSTE ONDERSCHEIDING
08:00
08:10
08:20
08:23.75
08:30
08:40
08:50
Iran neemt eerste nieuwe toestel Airbus in ontvangst
Amerikaans vice-president Joe Biden ontvangt Medal of Freedom
franse obama care obama_care zwitserse grens servië houdt verantwoordelijk
vliegtuigen iran president nieuwe airbus ontvangst genomen toestel landde vandaag teheran nucleaire deal opheffen sancties order tweehonderd plaatsen huidige vloot
autofabrikant onderscheiding krijgt grote staat die gediend streeft sterke middenklasse genoot vertrouwen fiat amerikaanse milieu
Metadata
Content Annotations

JOURNAAL; Journaal (2017): Laat Journaal

Title
JOURNAAL; Journaal (2017): Laat Journaal

Date
2017-01-12

Description
Met de volgende onderwerpen:

Index
daan-catalogue-aggr

ResourceId
2101712260240728631

Broadcaster
NOS

ExternalSourceInfo
View in DAAN (login required)

Persons
Weimans, Simone (presentatie)
Verhoef, Marco (presentatie)
Biden, Joseph (subject of discussion)
Clinton, Hillary (subject of discussion)
Zwart, Wouter (verslaggever)

Full metadata

Figure 1: Providing extra information using Linked Data while browsing data in the CLARIAH Media Suite

Finally, we evaluate our approach, reviewing the challenges and limitations. We discuss future work on both our linked data and its application in tools. For example, to support more use cases, to link to more collections, and to support users in more tasks that require combining data, such as analysis of persons based on characteristics such as gender and nationality.

Acknowledgement. This work was made possible by the CLARIAH-PLUS project funded by NWO (Grant 184.034.023).

KEYWORDS

digital humanities, cultural heritage, FAIR, linked data, interoperability, research infrastructure

REFERENCES

[1] S Münster, K Fritsche, H Richards-Rissetto, f Apollonio, B Aehnlich, V Schwartze, and R Smolarski. 2021. Teaching Digital Cultural Heritage and Digital Humanities the Current State and Prospects. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (2021).

[2] 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. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3 (2016).

[13] https://digital-skills-jobs.europa.eu/en/inspiration/research/digital-skills-cultural-heritage-report

[14] https://www.w3.org/TR/dx-prof-conneg/

[15] https://labs.beeldengeluid.nl/dataset/5520ccca-2c8e-11e6-a743-005056a71e3a