Optimizing Interoperability of Language Resources with the Upcoming IIIF AV Specifications

Jochen Graf
University of Cologne
jochen.graf@uni-koeln.de

Felix Rau
University of Cologne
f.rau@uni-koeln.de

Jonathan Blumtritt
University of Cologne
jonathan.blumtritt@uni-koeln.de

Abstract

In our presentation, we discuss how the upcoming IIIF AV specifications could contribute to interoperability of annotated language resources in the CLARIN infrastructure. After some short notes about IIIF, we provide a comparison between the concepts of the IIIF specifications and the ELAN annotation format. The final section introduces our experimental *Media API* that intends to optimize interoperability.

1 Introduction

The International Image Interoperability Framework (IIIF) (Snydman et al., 2015) is a technology agnostic standardisation for dissemination of web based images. It is driven forward by a large community of cultural heritage institutions. The original motivation behind IIIF is to optimize interoperability such that annotated image resources available at one institution can be reused by tools and services at other institutions. A side effect is that the framework facilitates implementation of web based image and annotation clients. With IIIF, clients can rely on well defined, feature rich, and stable application programming interfaces.

The IIIF Image API (Appleby et al., 2017a) and the IIIF Presentation API (Appleby et al., 2017b) build the core APIs of the framework. The IIIF Image API defines a set of low-level image manipulation requests, e.g., for image cropping, rotation, or format conversion. These low-level requests enable higher level features relevant for interoperability: for example, persistent web references not only to whole images but also to image details. The main idea behind the IIIF Presentation API is the so called *Canvas*[1]. *Canvas* represents a 2D coordinate space, where the target image(s) to be annotated and the annotations itself are organized together. The strength of the *Canvas* lies in its abstraction. The 2D canvas can be replaced by a canvas timeline for AV annotations with only small modifications on the specifications necessary.

Once an image-centric framework, IIIF is currently extended to other resource types from the cultural heritage domain, especially too for AV resources. Since 2016, the IIIF AV Technical Specification Group (IIIF AV Technical Specification Group, 2016) develops a new *AV Content API* mirroring the IIIF Image API in function and refines the IIIF Presentation API in order to make it equally useful for image, audio, and video annotation.

In the following, we aim to show how the upcoming IIIF AV specifications could contribute to the interoperability of language resources and to the development of web based AV annotation players - and the other way round: we aim to show that the ELAN annotation format forms an interesting case study to further develop the IIIF AV specifications.

2 ELAN IIIF AV Case Study

ELAN (Wittenburg et al., 2006) is a desktop annotation software for multimodality research. It is, among others, central to the research in the communities represented in the CLARIN-D working group *Linguistic Fieldwork, Ethnology, and Language Typology*. The tool produces time-aligned annotations in ELAN

[1] https://iiif.io/api/presentation/2.1/#canvas

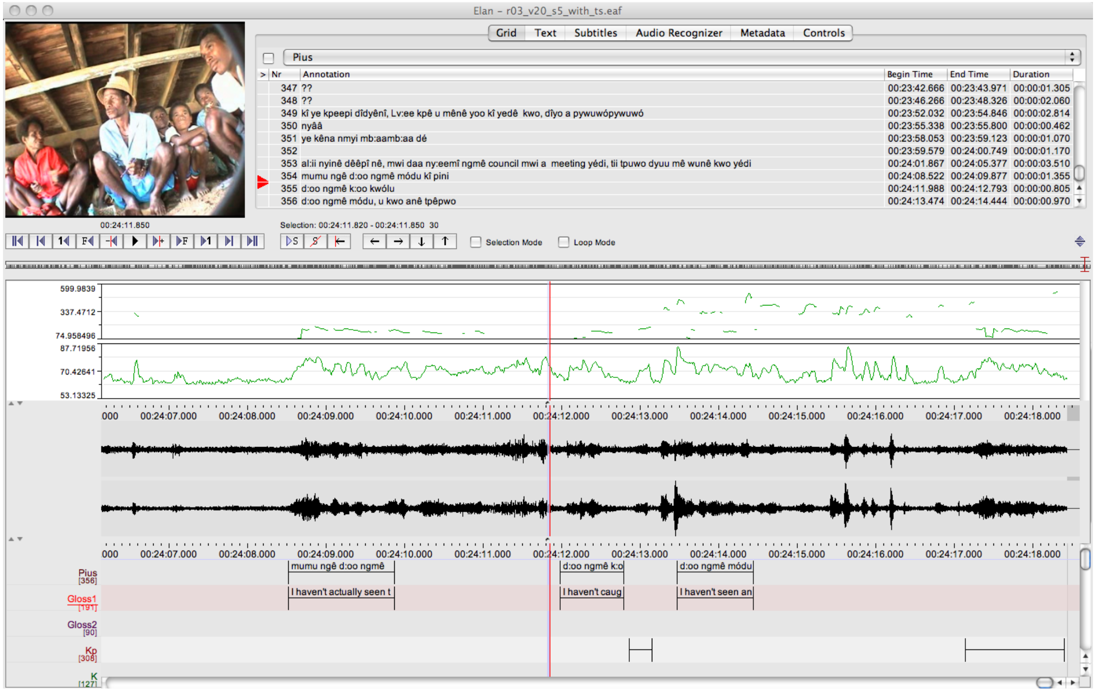

Figure 1: ELAN Main Window

Annotation Format (EAF) (Max Planck Institute for Psycholinguistics, 2006a), an open XML format. The ELAN tool is suitable for linguistic research since it does not only allow simple transcription of audio data but discipline-specific, time-aligned annotations up to the level of syllables and phonemes. A large number of EAF documents has become available in numerous language archives in recent years. To the best of our knowledge, there are only few archives that implement web based viewers for EAF Annotations in an adequate way (Berck and Russel, 2006)(Schroeter and Thieberger, 2006)(Sjölander and Beskow, 2000). As a result, the need for web interoperability of AV annotations should be as natural as for metadata (Freire et al., 2017).

Figure 1 (Max Planck Institute for Psycholinguistics, 2006b) shows the ELAN main window. Directly above and below the zoomable timeline in the middle, there are visual representations of the video's audio channels (intonation, waveform) - helpful tools providing researchers with an overview for an inherently time-bound and transient type of data. At the bottom, there are time aligned annotations grouped by layers shown in different colors. If one plays the referenced video file, the timeline and its attached intonation, waveform, and annotations are presented in form of a concurrent display with high time accuracy.

Having a deeper look into the ELAN annotation format, there appears much more complexity on the annotation level than visible on the user interface. The annotation format does not only support time aligned annotations[2] but also annotation references[3]. Annotation references have no direct timeline linkage but are linked to a parent annotation[4]. Additionally, there exist annotation types that can either subdivide the time range of a parent annotation having an own fixed start and end point in time[5], or types that dynamically divide parent annotations into a defined number of parts with equal length and without gaps[6].

When comparing the ELAN annotation format with the ongoing work done by the IIIF AV Technical

[2] <eaf:ALIGNABLE_ANNOTATION/>
[3] <eaf:REF_ANNOTATION/>
[4] Symbolic_Association
[5] Time_Subdivision and Symbolic_Subdivision
[6] Included_In

Specification Group, we identify overall accordance in respect to the way annotations are structured and grouped in lists and layers, enriched with different types of metadata, and linked to (parts of) media files. Those concepts can be easily mapped in both directions. We currently identify two differences, though:

Difference 1. The ongoing IIIF AV Content API specification does not yet propose the generation of visual representations of audio data such as spectrums or waveforms, although this is a useful utility for linguistic research. In the context of IIIF, it seems obvious to provide such visual representations in the form of images, respectively, with image tiles. A image tile of 500x25 pixels would contain the spectrum of a audio's time section with 10 seconds in length, for example. If a number of spectrum tiles is seamlessly strung together, concurrent display and deep zooming of time aligned annotations becomes possible in the browser as with ELAN, even for very large audio files.

Difference 2. There are many different types of annotations supported by the ELAN annotation format that, in their discipline-specific variety, seem to lay beyond the expressiveness of IIIF annotations. Since we could not find a convincing mapping, our current approach is to transform linguistic annotation references into standard, time aligned annotations accepting information loss. Since it is not our aim to provide a web version of the ELAN annotation tool, but only a player for presentation of AV annotations, the loss of information seems acceptable in regard to the increased interoperability.

3 Media API

3.1 Requirements Analysis

The requirements analysis for our experimental Media API started with a description of the *ELAN IIIF AV Case Study* in order to tie our experiment to a large real-world AV annotation dataset. Our Media API in any case should follow the IIIF AV specifications in the way that it mirrors the IIIF Image API in function: the proposed API should support all common transformations on AV media (cropping, format conversion, etc.) in a simple way as is the case for images and the IIIF Image API. Since our case study has shown that the scientific practice of linguistic annotation is well supported by visual representations of audio data, i.e., by spectrum or waveform image tiles, we like to propose that an ideal Media API would not only mirror the image API in function but would ideally be a superset of the IIIF Image API. In summary, we expect the Media API to cover the following function areas with at least the functionality mentioned in brackets:

Requirement A: common media transformations (format conversion, compression)

Requirement B: audio/video specific transformations (time cropping)

Requirement C: video/image specific transformations (cropping, scaling, rotating, color filtering)

Requirement D: audio to image transformations (spectrum and waveform extraction)

3.2 Derivation of the Media API from the IIIF Image API

The IIIF Image API defines five request parameters for image transformation as summarized below. According to the IIIF specifications, the parameters are processed in the order they are arranged in the URI from left to right: first, a rectangular portion of the input image is cropped, then the image is scaled, and so on.

Canonical URI Syntax of the IIIF Image API:

```
.../{region}/{size}/{rotation}/{quality}.{format}
```

Request Parameters of the IIIF Image API:

{region}	Defines the rectangular portion of the full image to be returned.
{size}	Determines the dimensions to which the extracted region is to be scaled.
{rotation}	Specifies mirroring and rotation.
{quality}	Determines whether the image is delivered in color, grayscale or black and white.
{format}	Format of the returned image.

Based on this API, we implemented our experimental *Media API* that allows to display linguistic annotations, visualizations of the audio signal as well as playback of the audio-visual data itself. The implementation prioritises interoperability with the IIIF AV specifications. For our API, we adopted the ongoing IIIF AV specifications and extended the concepts to fit time-aligned linguistic annotations.

Canonical URI Syntax of the Media API:

`.../{section}/{region}/{size}/{rotation}/{filter}/{quality}.{format}`

Request Parameters of the experimental Media API:

{section}	Defines the time portion of the full audio or video file to be returned.
{region}	Defines the rectangular portion of the full image or video to be returned.
{size}	Scales an image or video to a specific size.
{rotation}	Specifies mirroring and rotation for a image or video file.
{filter}	Applies filters to the input media file (waveform, spectrum, color, gray, bitonal, none).
{quality}	Defines the compression rate / quality scale of the returned media file (high, medium, low).
{format}	Format of the returned media file.

Our experimental media API, compared to the IIIF Image API, contains three AV related extensions:

Extension 1. In times of mobile devices used in low bandwidth networks, it seems desirable to offer audio and video data in different quality scales. For this purpose, we decided to reinterpret the $\{quality\}$ parameter directly before the $\{format\}$ ending due to its purely image related meaning (color, grayscale, black and white). $\{quality\}$ in our media API does not refer to the visual quality of the returned image but to the technical quality scale of the media file, where *high*, respectively *default* return the image, audio or video bitstream as is, *medium* and *low* return an increasingly compressed version of the input file with possibly human perceivable quality loss. Extension 1 fulfills requirement A.

Extension 2. The original *color*, *grayscale*, *black and white* functions are still there but are moved to the $\{filter\}$ parameter. Together with the $\{region\}/\{size\}/\{rotation\}$ URI part, requirement C is fulfilled and the API parameters together form a superset of the IIIF Image API. "Grayscale filter" or "bitonal filter" seem still acceptable names for the respective functions. The $\{filter\}$ parameter introduces flexibility for different media types: if the input file is a audio or video, one can apply a spectrum or waveform filter here. A spectrum image of 500x25 pixels in PNG format calculated from a 10 seconds audio section can be requested as follows:

`.../0,10/full/500,25/0/spectrum/default.png`

Extension 2 fulfills requirement D.

Extension 3. Finally, a $\{section\}$ parameter is put in front of all other parameters in order to allow cropping of time sections of audio and video files. This fulfills requirement B.

4 Conclusion

Adopting the ongoing IIIF AV specifications and extending its concepts to fit time-aligned linguistic annotations is showing promising results. Our current work concentrates on writing down a detailed technical documentation of our case study and on developing a prototype of our Media API - with the intent to report back our results to the IIIF AV Technical Specification Group. In order to achieve full interoperability within the CLARIN infrastructure, other issues have to be addressed, though: foremost, the issue of authentication and authorization of REST APIs in SAML-based authentication and authorization infrastructures needs further attention.

References

Michael Appleby, Tom Crane, Robert Sanderson, Jon Stroop, and Simeon Warner. 2017a. IIIF image API 2.1.1. https://iiif.io/api/image/2.1/. [Online; accessed 2019-04-09].

Michael Appleby, Tom Crane, Robert Sanderson, Jon Stroop, and Simeon Warner. 2017b. IIIF presentation API 2.1.1. https://iiif.io/api/presentation/2.1/. [Online; accessed 2019-04-09].

Peter Berck and Albert Russel. 2006. ANNEX a web-based framework for exploiting annotated media resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Nuno Freire, Glen Robson, John B. Howard, Hugo Manguinhas, and Antoine Isaac. 2017. Metadata aggregation: Assessing the application of IIIF and sitemaps within cultural heritage. In Jaap Kamps, Giannis Tsakonas, Yannis Manolopoulos, Lazaros Iliadis, and Ioannis Karydis, editors, *Research and Advanced Technology for Digital Libraries*, pages 220–232, Cham. Springer International Publishing.

IIIF AV Technical Specification Group. 2016. IIIF AV technical specification group folder. https://drive.google.com/drive/folders/0B8SS5OUXWs4GZ0ZfbEhIclhzb0k. [Online; accessed 2019-04-09].

Max Planck Institute for Psycholinguistics. 2006a. ELAN annotation format. http://www.mpi.nl/tools/elan/EAF_Annotation_Format_3.0_and_ELAN.pdf. [Online; accessed 2019-04-09].

Max Planck Institute for Psycholinguistics. 2006b. ELAN main window. https://tla.mpi.nl/tla-news/annex-and-elan-a-comparison. [Online; accessed 2019-04-09].

Ronald Schroeter and Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In *Sustainable Data from Digital Fieldwork. Proceedings of the conference held at the University of Sydney, 4-6 December 2006*. Sydney University Press.

Kåre Sjölander and Jonas Beskow. 2000. Wavesurfer an open source speech tool. In *Sixth International Conference on Spoken Language Processing*.

Stuart Snydman, Robert Sanderson, and Tom Cramer. 2015. The international image interoperability framework (IIIF): A community & technology approach for web-based images. *Archiving Conference*, 2015(1):16–21.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).