4. Data Format Control

System-level control over data formats is crucial for PECE's Open Knowledge mission (as stated in the introduction of this document) which comprises clear guidelines for generating, archiving, analyzing, and distributing Free and Open Source Software, Open Data, and Open Access publications. Data format control, in this respect, is tightly coupled in PECE's design with the perceived need for increased data accessibility and re-usability with interoperability among heterogeneous information systems.

In respect to its general guidelines for data format control and improved accessibility, PECE follows the Open Knowledge Foundation's Open Data definition observing three general principles for design and implementation of PECE's data management policies: 1) data must be discoverable and indexable through the web; 2) if the data is not machine-readable and distributed in an open format, it is not reusable; 3) open data must not not have legal restrictions for its usage, repurposing, and redistribution. For the purposes of data management, the PECE design team has adopted the OKF definition of "Open Knowledge" in working with the ethnographic data produced: "Open knowledge is what open data becomes when it's useful, usable and used" in the context of ethnographic projects.

In terms of technical specification, we described and implemented restrictions for content types and file formats that can be uploaded to the platform. The following table describes all the content types and the formats we use:

Content Type	Format	Extension	Commentary
Text	Hypertext Markup Language, Open Document Format, JavaScript Object Notation, Extensible Markup Language, JavaScript Object Notation for Linked Data, Resource Description Framework (UTF-8 encoded)	HTML, XML, JSON, JSON-LD, RDF, ODT, ODF	Serialized exchange file formats are delivered through the PECE Open API
Audio	OGG Vorbis, Opus, Advanced Audio Coding (Low Complexity), MPEG-1 Part 3, Microsoft WAVE Format 1	OGG, MP4, M4A, MP3, AAC, WAV (containers)	MPEG1 Part 3 (MP3), AAC, and WAV are proprietary technologies
Video	Theora, VP8, VP9, MPEG-4 Part 10 AVC (H.264)	OGG, OGV, WEBM, MPEG4, MP4 (containers)	MPEG4 Part 1 AVC and its MP4 container are proprietary technologies

Image	Joint Photographic Experts Group, Graphics Interchange Format, Portable Network Graphics, Scalable Vector Graphics	JPG, JPEG, GIF, SVG, PNG	
PDF document	Portable Document Format	PDF	

As the table demonstrates, we made an effort to adopt only “Web safe" and Open Document formats and standards. In doing so, we followed the guidelines of the W3C HTML5 standardization committee. There are, however, a few important exceptions to our Open format policy given the adoption of proprietary technologies (for containers and codecs of media files) as part of the W3C HTML5 specification. This is rather unfortunate given the state of dependency on proprietary video and audio technologies for the web. These exceptions include the adoption by the HTML5 video and audio tags with MPEG-4 part 10 AVC, as noted on the table above.

In terms of the actual implementation on PECE, data format control is executed at the interface level, that is, it is executed for data upload, presentation, and download. Through the web interface, only permitted formats are allowed to be uploaded. The user is presented with an error message when trying to upload a file that is not compliant with our Open format policy. After uploading a permitted file, we will use native support from web browsers that respect Open standards and formats (such as Mozilla Firefox, Chrome, Chromium, and Opera) to decode and render files on the browser (for all the supported formats: audio, video, texts, PDF documents, and images). For data harvesting purposes or for bulk operations, our Open API (as specified in the first section of this document on “Contextual Metadata Extraction”) operates with web standards for communication, authentication, and data manipulation and exchange (with JSON and XML formats).

In the roadmap for PECE 2.0 is the automatic, back-end transcoding of file formats: from proprietary and closed to open formats. We are testing and planning to implement audio and video transcoding capabilities on the platform as well as to offer automatic conversion of proprietary formats such as Microsoft Office Open XML to Open Document Formats, given their wider compatibility and sustained efforts to create interoperable, open, and community-driven formats.

Instructions for Developers: Media files depend on the installation of “video” module v.7.x-2.11 with the configuration of a video player based on “video.js” – whose dependency is the **“videojs”** module v. 7-x-2.3. We will not have video transcoding capabilities on PECE 1.0, we are already testing and evaluating the usage of server-side transcoding capabilities with FFmpeg on Drupal for PECE 2.0. PECE supports individual media files up to 700MB, so there is a need to configure PHP5 and Nginx with the parameters “upload_max_filesize” and

“post_max_size”. For file uploads, we have a listed of permitted formats for each content type: “PDF Document” (for PDF open standard files); “Video Artifact” (for `ogg`, `ogv`, `mp4`, `mpeg4`, `mkv`); “Audio Artifact” (for `ogg`, `aac`, `wav`, `mp4`, `m4a`); “Text Artifacts” (for `html text`, `odt`, `ods`, `xml`, `json`) with UTF-8 encoding; the other types of content on the platform include images and animations, which are restricted to web-safe formats: `jpg`, `jpeg`, `gif`, `svg`, `png`.