PROTOTYPING PHASE 2
FINAL REPORT

Project Acronym:	**PREFORMA**
Grant Agreement number:	**619568**
Project Title:	**PREservation FORMATs for culture information/e-archives**

MEDIACONCH

Revision: Final, October 31, 2016

Authors:

Dave Rice (MediaArea)
Jerome Martinez (MediaArea)
Ashley Blewer (MediaArea)

Dissemination Level		
P	Public	X

1 INTRODUCTION

During the PREFORMA Prototyping phase, suppliers are expected to provide software prototypes that fulfil the requirements of the PREFORMA project, to demonstrate the results of their development work, and to provide explanations and documentation (manuals) on how the developed software can effectively be used in archiving scenarios at memory institutions regardless of their size and the file type they make use of.

Following the same approach used last year, during the Second Prototyping Phase the plan for releases is as follows:

- Frequent releases: monthly;
- Intermediate releases: end of July 2016 and end of October 2016.

The intermediate release shall contain two parts:

- A functionally stable release
- A report which
    - Describes
        - In more detail the respective release;
        - The time line along with the current position (on time, delayed, ahead)
        - How suppliers managed to provide the required functionality (so far);
        - What is still missing compared to the original specifications and what is the plan to implement it.
    - Provides basic information to be used by PREFORMA WP8 in their deliverables to be submitted to the EC, reporting the work done by both suppliers and PREFORMA consortium members during the prototyping phase.

2 PROTOTYPING PHASE 2 - FINAL REPORT

1. Details
Type of Organisation: SARL
Registered Name of Organisation: MediaArea.net
Registered Address: Chemin du Vernay
Town/ City: Curienne
Postcode: 73190
County:
Country: France
Report Author: Jérôme Martinez
Telephone Number: +33 (0)6.12.65.98.37
E-mail Address: jerome@mediaarea.net
Project Name: MediaConch
Report Type: End of Phase 2 Report
Total Contract Price [euro]: 700 000 €
Start Date: April 14, 2015
End Date: December 31, 2016
Sub-contractors: Dave Rice, Ashley Blewer, Natalie Cadranel

1. Description of the release and progress compared to the last intermediate release

As with the previous report, MediaArea continues to help grow a community via the IETF (Internet Engineering Task Force) CELLAR (Codec Encoding for LossLess Archiving and Realtime transmission) working group, where work is being done to support the standardization of Matroska and FFV1. The working group's mailing list has been very active with members of the core Matroska community and the FFV1 community working on further developing the latest specification for standardization along with the MediaArea team. The listserv of the CELLAR working group now includes 114 members, including 48 active participants and over 950 emails.

MediaArea drastically overhauled the way policies could be created within MediaConch, including the ability for users to create a policy based on an existing video file report, the ability for users to compare two video files, and the ability for users to add complex logic to their policies.

The October release of MediaConch includes the first Metadata Fixer features, such as the ability to fix incorrect sizes within Matroska Elements. Much of our recent focus on the fixer had gone into a feature that uses the embedded CRC values of Matroska and FFV1 in order to correct minor amounts of damage. Now, in many cases, a Matroska or FFV1 encoding that experiences a "bit flip" can be corrected by MediaConch using the corresponding CRC as an error correction code. We've been coordinating with users and have additional fixer features prioritized for the upcoming 2016 releases. More information about fixing files with CRCs is available at https://mediaarea.net/MediaConch/fixity.html including examples on how this feature can be used to correct damage in FFV1 and PCM data.

A public policy page was added, permitting any user to share a policy with others. User can decide to share a policy by setting the "Policy visibility" to "Public" during the policy edition at https://mediaarea.net/MediaConchOnline/policyEditor. Public policies are available at https://mediaarea.net/MediaConchOnline/publicPolicies.

MediaConchOnline, the web interface, is now available without registration, but rather offers a guest mode. As our instance is public, we chose to protect our server by preventing each anonymous user to scan too many files, but this limitation can be removed in the configuration on another instance installed by someone else.

MediaConch also supports the newest versions of DPF Manager and VeraPDF in order to support analysis of TIFF and PDF in addition to Matroska and FFV1. MediaConch natively converts the reporting formats of DPF Manager and VeraPDF to support a consistency in reporting and display (via XML or HTML or text outputs). MediaConchOnline is configured with versions of DPF Manager and VeraPDF (based on September releases) and can be tested at https://mediaarea.net/MediaConchOnline.

As part of a collaboration with VIAA, MediaArea used MediaConch's plugin abilities to integrate an FFmpeg plugin. This allows MediaConch to be used in order to convert preservation files to Matroska/FFV1 files while running the implementation checker. We also continued to improve the REST API and now the web version of MediaConch uses the REST API for all the features. Additionally, MediaConch has continued to establish presence in a variety of communities that have interest in the developing software by presenting or being represented at many different conferences.

Progress Highlights Since the July, 2016 Release

Since the July, 2016 release, the project has made significant progress towards prior objectives and more recent goals arising from our user experience research. These include: upgrades to policy checker complexity, integration into Artefactual's Archivematica, conference presentations, hosting a symposium, webinars, community outreach, and new, continued collaboration with preservationists and developers within the IETF CELLAR working group.

- Major overhaul of policy checker
- Initial Metadata Fixer features and data correction features
- Updates to policy and implementation report displays
- Integration of VeraPDF and DPF Manager into MediaConch
- Conference presentation at IASA
- Presentation on CELLAR at iPRES
- Workshop on MediaConch at iPRES
- Webinar with PREFORMA
- Webinar with Artefactual
- Workshop at Tate with PERCILES project

In the upcoming months, MediaArea expects to increase the count of system policies (default policies provided and supported by us), improve the FFV1 specifications, add more features to the fixer module, add more feature to the public policies feature, stabilize the software, and make updates based on new requests from the users on the project issue tracker.

2. Datasets used to test the release
The MediaArea team helped moved the official Matroska collection of test files from a static zip file hosted on sourceforge into a new GitHub repository at https://github.com/Matroska-Org/matroska-test-files. In addition, MediaArea has re-processed a collection of a hundred thousand original Matroska files at archive.org under new more verbose parsing options of MediaInfoLib in order to generate more comprehensive test results in the application of the developing the implementation checker and to re-assess our checker. The original work to analyze the massive corpus at the Internet Archive has been redone, since the original analysis parsed files too selectively to cover many types of Matroska implementation checks. The new analysis is more comprehensive and allows MediaArea and the CELLAR working group to identify implementation errors and relate them to specific samples, specification language, and software. Additionally, MediaArea has created a process to generate FFV1 samples under the various options allowed under historical FFmpeg builds. The use of these datasets and the development work in CELLAR has also clarified many obligations, constraints, and expectation for Matroska validity so the need for comprehensive test sets has greatly expanded. Testing with the policy checker has expanded in our collaborations with Artefactual and the Tate Museum. Through testing and collaboration with these organizations and other users, we've determined that the policy checker had to be significantly expanded to allow for more complex and conditional rules as well as allowing the expression of policy for values between two provided files. For instance, Archivematica could have a policy that states that an original file and created derivative should share specific significant characteristics or the Tate may use complex and conditional policies to test if media adheres to the requirements of various display hardware. The policy work with the Tate led to an expansion of policy logic to include reporting data from MediaTrace in addition to our existing use of MediaInfo for very refined structural tests. Additionally, when working with our collaborators, we found the original policy logic too simple to support many in-practice policies which required conditional logic. A blog post about this is at https://mediaarea.net/MediaConch/2016/10/04/policy-refactor/. In the No Time to Wait symposium, the MediaArea team curated a set of presentations that focused on testing FFV1, included Peter Bubestinger's work to test performance and resilience of FFV1 under its massive combinations of options. Additionally Kieran Kuhnya presented on his work to create a damage and decode about 22 million FFV1 files in order to isolate decoding errors, file tickets, and research that has helped refine the FFV1 decoder and clarify some requirements with the specification language.

3. Dissemination and community building
- Dave Rice and Tessa Fallon gave an update to the IETF CELLAR working group and FFV1 and Matroska standardization at IASA in Washington, D.C., USA September 27th. - MediaArea wrote a whitepaper on the standardization process for FFV1 and Matroska entitled "Status of CELLAR: Update from an IETF Working Group for Matroska and FFV1." This paper is available in the proceedings here:http://www.ipres2016.ch/frontend/index.php?page_id=3276 - Ashley Blewer presented on this work at the iPRES conference in Bern, Switzerland October 4th. Programme: http://www.ipres2016.ch/frontend/index.php?folder_id=353 - MediaArea also participated in a PREFORMA workshop on October 5th, demonstrating the MediaConch tool and how it can be applied to digital preservation workflows. Ashley Blewer and Jerome Martinez held a breakout session where future users can ask questions. - CELLAR continues to grow in number. On October 30th, MediaConch team members Ashley Blewer and Dave Rice held an informal meetup with Nick Krabbenhoeft at New York Public Library to introduce interested CELLAR members to the necessary work. - Artefactual continues to integrate MediaConch into Archivematica, as summarized in the previous report. - On October 18th, Dave Rice and Natalie Cadranel hosted a webinar with Artefactual on their integration of MediaConch. Video: https://www.youtube.com/watch?v=ZTG9nlp_4oA - Artefactual will give a poster presentation at the Association of Moving Image Archivists conference in early November regarding Archivematica and MediaConch. http://www.amiaconference.net/preliminary-program-2/ - Dave Rice will teach usage of MediaConch as part of a "Digipres 101" workshop held at the Association of Moving Image Archivists. http://www.amiaconference.net/pre-conference-workshops-symposia/ - Dave Rice and Ashley Blewer gave two workshops at the Tate in late July. Here is the Tate and PERCILES project summary report: A workshop on the use of Mediainfo, MediaConch and FFmpeg in the preservation of digital video: http://pericles-project.eu/blog/post/TateWorkshop2016 - MediaInfo (Jérôme Martinez) and FIMS had an EBU booth at IBC (EBU booth), including discussion on MediaConch features, on September 10th - MediaConch cited in AV Digitisation and Digital Preservation TechWatch Report #04 : https://www.prestocentre.org/library/resources/av-digitisation-and-digital-preservation-techwatch-report-04. - Plans to expand software related to community-building will include ability for users to share their policies and a wiki for embellishing failed implementation reporting, which involves the ability for users to contribute more thorough details about file format failures. - MediaConch is now part of the latest official Ubuntu version (MediaConch is directly provided by Ubuntu on their DVD, no need to any external resource) and part of Debian SID (the “work in progress” version, which will be the next official Debian version), as well as Homebrew (open source repository) in order to meet the classic requirement for good, open source community building. We are working with other Linux distribution (e.g. Fedora, Arch...) in order to reach the different open source communities.

4. Open Source approach

MediaArea continues to work closely with FFV1 and Matroska format designers to clarify the format specifications. This work is done in the open via the IETF CELLAR listserv and Matroska and FFV1 Github pages so that it is transparent and anyone can contribute on each issue. The Matroska specification page hosted by MatroskaOrg in GitHub and now contains 136 commits by 7 code contributors (and many more commenting or giving feedback on Github or via CELLAR listserv communication). The foundational format of Matroska, EBML, has additionally received 296 commits from 7 contributors and support from the working group members.

MediaArea encourages community members to get involved through 1-on-1 coaching and instructions, speaking at conferences, and holding workshops about MediaConch and, more generally, about file validation and conformance checking. In November, MediaArea plans to co-host an unofficial event to support learning about and assisting in the standardization of Matroska and FFV1 at the New York Public Library. Both MediaConch and Archivematica (who has integrated MediaConch) are leading a workshop in training at the Association of Moving Image Archivists to show archivists how to participate, contribute to, and use these open source projects.

MediaArea also funded development of FFV1 and Matroska features in external open source projects in support of the goals of the PREFORMA Challenge. Through these efforts we offered financial support to several development tickets. Notably the EBML specification (the underlying format of Matroska) had always said that Top Level Elements of Matroska should store CRC checksums of the data stored within the element. MediaArea funded the development work and as of FFmpeg 3.2, Matroska files are written with the CRC and fixity recommendations of the specification, enabling the data to have documented fixity from the initial writing. By running massive and diverse Matroska collections through MediaConch, we supplied data back to open source projects that wrote such files supporting many bug fixes, optimization, and improvements.

We also contributed support to FFmpeg developers to improve the effectiveness for FFV1 in film preservation and contributed back recommendations and reviews to the FFmpeg encoder and decoder of FFV1 as issues were discovered. For example, for the 16-bit integer overflow issue, we decided to update the spec for handling a special case for YUV 16-bit as there were already such encodings in the use, but we kept the generic algorithm for RGB 16-bit as there was not yet any encoder supporting this encoding.

A provision of the source code made by MediaArea is to conform to the agreement between the 6 suppliers from the first phase and the PREFORMA consortium: as stipulated in the “Clarification from PREFORMA on licensing requirement” sent by Claudio Prandoni February 26, 2015, in an email having the subject “Re: [PREFORMA OS Projects] Response from suppliers on licensing requirements”, we provide in our deliveries to PREFORMA this way (sentences taken from PREFORMA answer): “All code (software and libraries) required to compile and/or execute the Conformance Checker in a production environment has to be freely available in open source form under generally recognized free software licenses compatible with the GPLv3++ and MPLv2++ to enable redistribution of the whole package under these two licenses.”

For this reason, we provide some code e.g. sha2.c or tinyxml2.cpp under their respective licenses (BSD 3-clause or zlib licenses), and we depend on external libraries, not GPLv3++ and MPLv2++, on purpose. The reasons (summary: not reinvent the wheel) were explained in the initial email sent by Jérôme Martinez in the name of the 6 suppliers on the email called “Response from suppliers on licensing requirements” on February 19, 2015.

MediaArea took careful consideration of patent issues and has acted to significantly reduce (hopefully remove) patent risk in the code delivered to PREFORMA. For example, “MPEG-4 and JPEG 2000” are not decoded as streams but only frame headers are analyzed.

In our aim to meet best practices of open source, we reuse components instead of coding new tools, e.g. we use Qt open source UI toolkit for displaying the local UI. As a result, the size of the delivery may be considered to be large on platforms not having the open source approach (e.g. Windows and Mac), however the size is not so big as the command line version of MediaConch is 4 MB in size including all necessary dependencies (no need to download an external package) for Linux, Mac and Windows, and the Graphical version of MediaConch is 3 MB in size for Linux and 25 MB in size for Mac and Windows. If PREFORMA chooses to support only Matroska/FFV1/PCM, it is possible to configure MediaConch without the support of other formats potentially used by archives (e.g. MOV, AVI, MXF…) in order to reduce the size of the binaries.

5. Standardisation efforts

MediaArea continues to work actively on standardization efforts in several directions. Our team has actively contributed to and brought communities around the standardization efforts of EBML, Matroska and FFV1. Since our July report, the draft for EBML has been upgraded to a working group items and received another revision in the CELLAR document tracker. Within the EBML specification, we helped draft the concept of the EBML Schema, which defines an EBML Document (of which Matroska is one type) in a schema which can be used to validate a document similar to how an XML Schema can function. This effort helps facilitate and consolidate many implementation checker tests as the specification for Matroska now has a more machine-readable form.

Also in the CELLAR working group, we have aided in rewriting sections with clarified language and contributed recommendations from mining our Matroska datasets. We also adapted a draft of Google's draft of 360 degree and virtual reality elements for Matroska presentations to held aid in keeping Google's work in webm and CELLAR's work in Matroska well aligned. We look forward to continuing as participants in these efforts as more in preservation communities seek to use these formats.

In order to facilitate the implementation of these standards we have sponsored and participated in patches to related open source projects (see the Open Source section), notably supported the integration of CRC and fixity features into FFmpeg's Matroska muxer. Recent FFmpeg builds include many new features tied to recent CELLAR work such as support for undefined aspect ratios, more accurate handling of aspect ratios, better description of color properties, and better interlacement support. Additionally the FFV1 encoder and decoder have also been extended with new features defined by CELLAR.

We hope to soon support a training seminar to help teach interested standardization participants in the tools of the working group, such as Markdown and GitHub. The Association of Moving Image Archivists occurs in November and members of our team will lead workshops and meetings related to inclusiveness and accessibility of the standardization efforts.

Within the MediaConch project itself, we have developed XML Schemas for each of our XML based reports to better define standardize the XML-based expressions used in the software.

We have also continued in community work to prepare for a set or recommendations for the archival use of Matroska and FFV1 in a manner analogous to the Library of Congress's similar standardization effort in AS-07 which relies upon MXF and JPEG2000.

6. Impact assessment, sustainability, future use and exploitation

Extending into communities via Archivematica

MediaArea's partnership with Artefactual includes MediaConch being integrated into Archivematica workflows. Archivematica is a popular framework for OAIS-compliant digital preservation with a robust user community. Initial integration of MediaConch is available in the latest Archivematica release, which has been promoted during Artefactual's Archivematica Camp and a webinar (Video: https://www.youtube.com/watch?v=ZTG9nlp_4oA) specifically on the topic. Artefactual will give a poster presentation at the Association of Moving Image Archivists conference in November on MediaConch and the integration and use within Archivematica.

MediaConch integration at VIAA and Workshops at Tate Museum

MediaConch has collaborated with VIAA and PACKED in order to integrate MediaConch into VIAA's archival environment. This included utilizing the plugin architecture of MediaConch to add transcoding support via FFmpeg into MediaConch. Thus MediaConch will be use to transcode archival videos into Matroska/FFV1, use the policy checker to assure that the source file and resulting Matroska/FFV1 share significant characteristics, and run the implementation checker on the result.

MediaArea has collaborated with the Tate Museum in the context of their Pericles project in order to teach two workshops and extend the MediaConch policy checker to include methods to develop policies that compare two files to each other (for example to compare a source file and result after transcoding to ensure that only certain characteristics are adjusted). This collaboration also led to more advanced policy checking features in order to test certain files for specific hardware support.

Extending into other formats

Functionally, MediaConch can already be expanded to support file formats beyond Matroska, FFV1, and LPCM (and PDF/TIFF via integration of work from the other suppliers). There is potential for MediaConch to become the conformance checking software for *any* audiovisual formats after the completion of this project, not just limited to Matroska and FFV1. MediaConch's local policy creation feature can already be extended out to any format that MediaInfo supports. For this reason, MediaConch is delivered with some basic support of other formats used by archives e.g MOV, AVI or MXF, as a demonstration that MediaConch is versatile. MediaArea has promising projects coming in 2017 to help us further sustain, extend, and promote MediaConch.

Future use

MediaArea has started planning for the commission of supplemental media format diagnosis and support to be added to MediaConch after the project ends by beginning conversations with cultural heritage institutions and gauging interest in supplemental media format sponsorship and software integration opportunities for institutional workflows.

7. Gap analysis and next steps

We have been behind schedule regarding the metadata fixer, opting instead to prioritize the "performance optimization" of the application, increase implementation checker work, and an overhaul of the policy checking elements of MediaConch, as well as prioritizing features not in our initial plan but requested by our users like a Public Policies page and an FFmpeg plugin.

However the October release of MediaConch features our initial fixer, which starts with a key fixer feature that uses the embedded CRC's of Matroska and FFV1 to fix flipped bits and correct these if found, saving the file from damage. While progress was behind schedule for the fixer, the work will be implemented in this release.

The development of the Implementation Checker has been an ongoing chase against the progress of the Matroska and FFV1 specifications within the CELLAR working group. As the initial work has calmed, the implementation checker has become more stable and refined. The development of the EBML Schema by CELLAR has enabled MediaConch to use an EBML Schema similar to how an XML Schema is used for validation. However more changes, clarifications and refinements are expected from CELLAR and we plan to continue developing MediaConch to follow and contribute to CELLAR for the betterment of both projects. As regions of the standard are clarified, implementation checks are added to MediaConch.

Our growing community has expressed the need for a place to share policies, which we will implement in the WebUI in the beginning of November as a collaborative web framework for sharing institutional policies.

Verbosity settings are present in MediaConch but should still continue to be expanded to be more complex. Current verbosity is "high" and "low" and lack nuanced versions in between. There are plans to extend verbosity to 5 distinct levels.

We have found that more opportunities exist to optimize the implementation checker. The checking process must often cover tens or hundreds of thousands of Matroska elements and FFV1 frames per file, so optimization is essential to expanding use and community.

LPCM has not been given adequate attention in comparison to the other two formats under commission, FFV1 and Matroska. LPCM, because it is not going through the standardization process and is relatively simple, has been the last to receive implementation checks in MediaConch. This has been complex since nearly any data stream can be considered as a valid LPCM stream. The introduction of the CRC features in the MediaConch fixer allow some damaged PCM in Matroska (with CRC Elements) to be corrected.

For further context of our current tasks and next steps please see our issue trackers at https://github.com/MediaArea/MediaConch_SourceCode/issues (interfaces), https://github.com/MediaArea/MediaInfoLib/issues (analysis of files), and https://github.com/MediaArea/MediaConch/issues (documentation).