DPF Manager

The open source
COMMUNITY

DPF MANAGER

PHASE 1 CLARIFICATIONS AND UPDATES

Project acronym: PREFORMA
PREFORMA - Future Memory Standards
PREFORMA - Future Memory Standards
PREservation FORMATs for culture information/e-archives
EC Grant agreement no: 619568

INDEX

1. Introduction

Thanks to the questions we were asked during the 6th of March presentation, and the feedback we received from the PREFORMA members after the presentation was finished, we were able to identify areas of our proposal that were not as clear or as strong as we thought they should be.

One of the technical questions we were asked was about the use of 3rd party libraries in the DPF Manager, and how were we planning to deal with color profiles.

To provide further clarification on both topics, we have attached as an annex the document "Architecture clarification document", which we have also merged with the Technical specifications document (also submitted as version 2).

We were also asked about our experience managing open source projects and communities. This is probably not one of the strongest points in our proposal, so we decided we would benefit from bringing an expert into our team to lead that part of the work. In order to select the best candidate, we have received the help of a member of the Free Software Foundation, who has used his network of contacts to help us find and evaluate the potential candidates for the job.

After thorough consideration we have selected Bas Van Leeuwen as the person who will lead the open source efforts of the DPF Manager. Bas is the General Secretary for the Dutch Community Managers Association, and has experience managing open source communities with 500+ developers.

Attached as an annex is Bas CV, and the added section to the document "EASY INNOVA - 6. Open Source Best Practices and Dissemination Plan.pdf" (also submitted as version 2).

Finally, we are aware that as an SME we would not have easy access to the biggest memory institutions as they are used to working with leading companies in the field.

We started looking at big players in the industry that we could partner with. Thanks to being located in the same technology park, we were able to present the DPF Manager to Hewlett-Packard CDS, which were really interested in a future commercial partnership. They also showed interest on a possible integration of the DPF Manager functionality into HP MediaBin (Hewlett-Packard digital asset management solution) and other DAM software.

Attached as an annex is a letter of intent from Hewlett-Packard CDS, and the added section to the document "EASY INNOVA - 7. Business Commercialization Plan.pdf" (also submitted as version 2).

2. List of annexes

Annex A contains the changes made to the document “EASY INNOVA - 7. Business Commercialisation Plan”

Annex B contains the letter of intent from Hewlett-Packard CDS

Annex C contains the changes made to the document “EASY INNOVA - 6. Open Source Best Practices and Dissemination Plan”

Annex D contains the CV of Bas van Leeuwen

Annex E contains the changes made to the document “EASY INNOVA - 3 Technical specifications”

DPF MANAGER

ANNEX A: CHANGES TO THE BUSINESS COMMERCIALIZATION PLAN

Project acronym: PREFORMA
PREFORMA - Future Memory Standards
PREservation FORMATs for culture information/e-archives
EC Grant agreement no: 619568
EC Call ID: FP7-ICT-2013-11

Changes made to the document “EASY INNOVA - 7. Business Commercialization Plan”: added new section 2.3

2.3 Partnerships with big players in the industry

It is not always easy for SMEs to have access to big memory institutions as they are used to working with well-known and respected big players in the industry.

In order to overcome this issue, we have started contacting big companies that we have identified could help Easy Innova with the integration with legacy software, commercialisation and support for DPF Manager.

So far we have entered into talks and have a letter of intent from **Hewlett-Packard CDS**, a Hewlett-Packard subsidiary dedicated to the integration and support of HP and third party software worldwide. We have presented the DPF Manager to them, and we have agreed to explore the possibility of integrating DPF Manager functionality into **HP MediaBin** (Hewlett-Packard digital asset management solution) and other DAM software. This contact has been possible because our offices are located in the same technology park as Hewlett-Packard CDS office in Girona.

Based on the questionnaire we sent to numerous memory institutions asking what software they use to manage they archives (amongst other things), we have identified the following vendors that already have an established reputation inside the digital preservation community:

Company	**Software solution**
ExLibris	Rosetta
DuraSpace	DSPACE
Canto	Canto Cumulus
Tessella	Tessella SDB
Libnova	Libsafe
Axiell	Axiell CALM
Preservica	Preservica Cloud, Standard and Enterprise edition

Therefore, they are important candidates to be contacted and to create a commercial partnership if possible. Of course, we will also keep adding companies to the list as we learn about them. The fact that we already have support from Hewlett-Packard CDS will make it easier to establish a working relationship with these.

DPF MANAGER

ANNEX B: LETTER OF INTENT OF HEWLETT-PACKARD CDS

Project acronym: PREFORMA

PREFORMA - Future Memory Standards

PREservation FORMATs for culture information/e-archives

EC Grant agreement no: 619568

EC Call ID: FP7-ICT-2013-11

CDS
An HP Company

Hewlett-Packard
Customer Delivery Services, S.L.
Parque Empresarial Las Rozas
C/ José Echegaray, 8 Edificio 1
28232 Las Rozas (Madrid)

Dr. Miquel Montaner
Chief Technology Officer
EASY INNOVA SL
Parc Científic i Tecnològic de la UdG
C/ Emili Grahit, 91 Ed. Narcís Monturiol
17003 Girona
Spain

March 12th, 2015

Dear Dr. Montaner,

As you know Hewlett-Packard CDS is a service-oriented business with a strong focus on quality of service and customer satisfaction.

After your presentation and looking at the documentation and business plan of DPF Manager we would like to express our interest in the project. We believe that there are potential sources of revenue that are worth exploring together.

We think that your proposal to establish a commercial partnership between Easy Innova and Hewlett-Packard CDS merits further discussion, and we will follow with interest the development of the project.

Sincerely,

Miquel Tarragona,
Service & Delivery manager
Hewlett-Packard CDS

hp

DPF MANAGER

ANNEX C: CHANGES TO THE OPEN SOURCE BEST PRACTICES AND DISSEMINATION PLAN

Project acronym: PREFORMA

PREFORMA - Future Memory Standards

PREservation FORMATs for culture information/e-archives

EC Grant agreement no: 619568

EC Call ID: FP7-ICT-2013-11

Changes made to the document “EASY INNOVA - 6. Open Source Best Practices and Dissemination Plan”: added new section 1.2

1.2 Leadership in the creation of the community

We have sought to bring an international expert in the creation and growth of online communities around open source software products into the team, as this is an area we feel we could benefit from external expertise.

During the selection process we have received the assistance of a member of the Free Software Foundation, who has used his network of contacts to help us find and evaluate the potential candidates for the job.

After thorough consideration we have selected Bas Van Leeuwen as the person who will lead the open source efforts of the DPF Manager. Bas is the General Secretary for the Dutch Community Managers Association, and has experience managing open source communities with 500+ developers.

Bas will lead the open source efforts for the DPF Manager.

DPF MANAGER

ANNEX D: CURRICULUM VITAE OF BAS VAN LEEUWEN

Project acronym: PREFORMA
PREFORMA - Future Memory Standards
PREservation FORMATs for culture information/e-archives
EC Grant agreement no: 619568
EC Call ID: FP7-ICT-2013-11

Bas van Leeuwen

Principal Consultant at Evidently Community

Amsterdam Area, Netherlands | Computer Software

Current	Evidently Community, Community Managers NL
Previous	Hippo, Universiteit van Amsterdam
Education	University of Amsterdam

Background

Experience

Principal Consultant
Evidently Community

February 2015 – Present (2 months) | Amsterdam

Available for advice, coaching, training, and hands-on community management

General Secretary
Community Managers NL

January 2015 – Present (3 months) | Netherlands

General secretary of the Dutch Community Management association; responsible for the day-to-day goings on of the association.

Community Manager
Hippo

January 2013 – January 2015 (2 years 1 month)

I am responsible for engaging and maintaining the various Hippo Communities (developers, users). Engaging users both online and offline at various meetups and Hippo's annual GetTogether (hgt14.onehippo.org) that last year saw 175 people from all over the globe.

Business Consultant / Project Manager
Hippo

September 2009 – December 2012 (3 years 4 months) | Amsterdam Area, Netherlands

I helped clients achieve the best solution for their web-needs. The variety that Hippo CMS provides gave me the opportunity to be creative and (together with key developers) devise the best result. Implementations were either done by in-house development teams, or by Hippo partners, both of which I supported where needed. Also helped out with Hippo's community events, both in-house (Hippo Fridays) and external (Hippo GetTogethers).
I did various courses and am certified IPMA-D.

Example projects:
- Launch of Politie.nl
- NS.nl restructuring
- Amerpoort.nl

Support Engineer
Hippo

October 2008 – May 2009 (8 months)

Managing client's issues using Hippo's JIRA system, providing feedback to clients and helping Hippo fulfill it's service level agreements.

Student advisor
Universiteit van Amsterdam

September 2006 – June 2007 (10 months)

Tasked with advising future students with their choice of education.

Volunteer Experience & Causes

Vice president
Vereniging Informatiewetenschappen Amsterdam
October 2005 – October 2006 (1 year 1 month) | Education

As vice president I helped the president in his duties, took care of sponsorship and replaced the president where needed.

Chief editor online
Cl!max - Schoolkrant Montessori Lyceum
January 2003 – August 2004 (1 year 8 months) | Education

Responsible for the forums, the site and the web editors.

Languages

English
Full professional proficiency

Dutch
Native or bilingual proficiency

French
Elementary proficiency

German
Limited working proficiency

Chinese
Elementary proficiency

Education

University of Amsterdam
B.Sc., Computer Science
2004 – 2009

Shanghai International Studies University
Beginner course in Chinese
2007 – 2008

A semester of Chinese language and full immersion in the Shanghai city life.

DPF Manager

The open source
COMMUNITY

DPF MANAGER

ANNEX E: ARCHITECTURE CLARIFICATION DOCUMENT

Project acronym: PREFORMA
PREFORMA - Future Memory Standards
PREservation FORMATs for culture information/e-archives
EC Grant agreement no: 619568
EC Call ID: FP7-ICT-2013-11

INDEX

1. Document Overview

1.1. Purpose

The Purpose of document is to provide clarification to some of the technical questions raised around the DPF Manager architecture in the 6th march presentation. This document contains a more extensive description about the 3rd party libraries listed in the document *"EASY INNOVA - Technical specifications.pdf"*. It also includes additional documentation about the conformance checker internal architecture design without using any external library dependency.

The clarifications in this document have been added to the technical specifications document, in sections "4.3.1 Conformance Core Module" and "7.4 Use of third party libraries".

1.2. Document structure

Section 1 provides the purpose, the structure of this document as well as references used in this document.

Section 2 provides an introduction to the document.

Section 3 describes the 3rd party libraries we are planning to use in the DPF Manager and clarifies where they are going to be used and why. It also describes how we will implement a Conformance Checker without using external dependencies.

1.3. Document References

This Section provides a list of the references that were used in preparation of this document:

TIFF Revision 6.0 Final — June 3, 1992

Adobe Photoshop® TIFF Technical Notes March 22, 2002

Adobe PageMaker® 6.0 TIFF Technical Notes September 14, 1995

EASY INNOVA – 3. Technical specifications.pdf

EASY INNOVA - 5. Intellectual Property Rights report.pdf

ICC.1:2010 Image technology colour managements – Architecture, profile format, and data structure.

ISO 12234-2:2001 Electronic still-picture imaging -- Removable memory -- Part 2: TIFF/EP image data format

ISO 12639:2004 Graphic technology -- Prepress digital data exchange -- Tag image file format for image technology (TIFF/IT)

XMP specification part 1 - Data model, serializations and core properties. April, 2012

XMP specification part 2 - Additional properties. April, 2012

XMP specification part 3 - Storage in files. May 2013

Exchangeable image file format for digital still cameras: EXIF Version 2.3. December, 2012

Dublin Core Metadata Initiative, August 2007

IPTC Standards - Photo Metadata White Paper, 2007 Document Revision 11

PREMIS Data Dictionary for Preservation Metadata, July 2012

METS METADATA ENCODING AND TRANSMISSION STANDARD: PRIMER AND REFERENCE MANUAL , 2010 Digital Library Federation

2. Introduction

This document pretend to be a clarification and further detailed description of the architecture design presented in the 6th march deliverable and in anyway pretend to include any architecture modification in the previous documentation provided.

The Architecture proposed by EASY INNOVA has been designed to be developed from scratch. Aside from the 3rd party libraries listed in the technical documentation, we do not want to reuse any exiting code already developed for EASYINNOVA in other projects and for other purposes.

The aim of following this rule is:

- Avoiding compromising any of our design principles defined in "EASY INNOVA - 3. Technical specifications.pdf, section 3.2, page number 9" trying to fill already developed software in a new architecture.
- Developers do not have to learn other frameworks or libraries that will increase the DPF manager framework learning curve.
- No need to sublicense any existing software.
- The lack of innovation that could involve using already developed software.

We can overcome the lack of libraries in the core application working in the following points.

- **Domain expertise:** Library authors are usually experts in the domain covered by the library. We have deeply studied all the standards involved in the Conformance Checker, and we are working together with experts in image preservation from the University of Basel in order to develop the DPF Manager Framework.

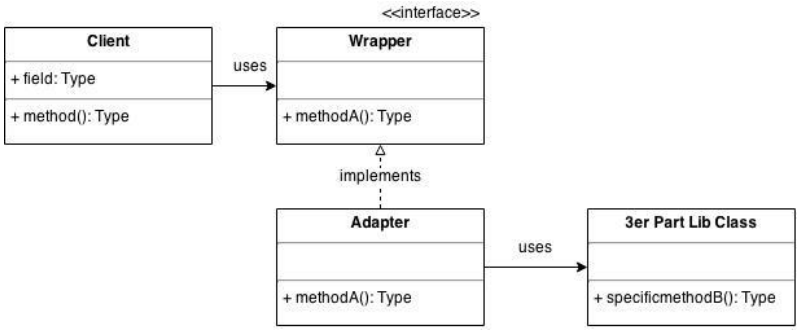
- **Stability:** 3rd party libraries are being used by other people as well as you and in many cases, hundreds if not many thousands of developers worldwide. Most of the early problems have already been detected by others and fixed by authors. We hope that our developer's community will give us this stability in the future but meanwhile we will perform unitary test and use of integration platforms to counterbalance.

- **Financial impact:** We have evaluated the viability in cost and time in terms of developing our Conformance Checker from scratch to ensure that is completely possible.

There are some points in the architecture where the use of 3rd party libraries is mostly required. We have evaluated different frameworks and we have chosen the libraries taking into account; libraries already used successfully in other projects, widely used, stable, tested, active community, bugs replaced and documented.

In the Architecture design we also take into account the possibility to replace a currently used library for another one using the adapter pattern.

With the Adapter pattern we could wrap the third party library class with and interface (Wrapper) that we expose to the developers. Developers will directly use the exposed interface and not have to directly call third party library methods. This provides a way to shield or break interfaces when a third party library changes something. To use another library you only have to implement a new Adapter.



In the following sections we explain where the libraries selected are going to be used as well as why there are used.

All the libraries are compatible with GPL v3+ and MPL v2+ as well as their dependencies. This compatibility has been approved by Malcolm Bain, our legal advisor, a report is attached in the documentation sent in the March 3rd deliverable "EASY INNOVA - 5. Intellectual Property Rights report.pdf".

-

3. Architecture Clarifications

3.1. Use of 3r party libraries

3.1.1. General libraries

As explained in the "EASY INNOVA - 3. Technical specifications.pdf" section 7, page 58 the programing language selected for the DPF manager is Java, using the version 7. All Java versions include a library known as JDK. This base library distributed under GPLv2+CE is suitable for most cases but have some missing functionalities.

Apache commons is one of the most widely used libraries in Java. This library implements some of common missing functionalities in the JDK or extends the existing ones. Apache Common components do not have dependencies on other libraries, so that these components can be deployed easily. The Apache Common components that are going to be use are the following ones.

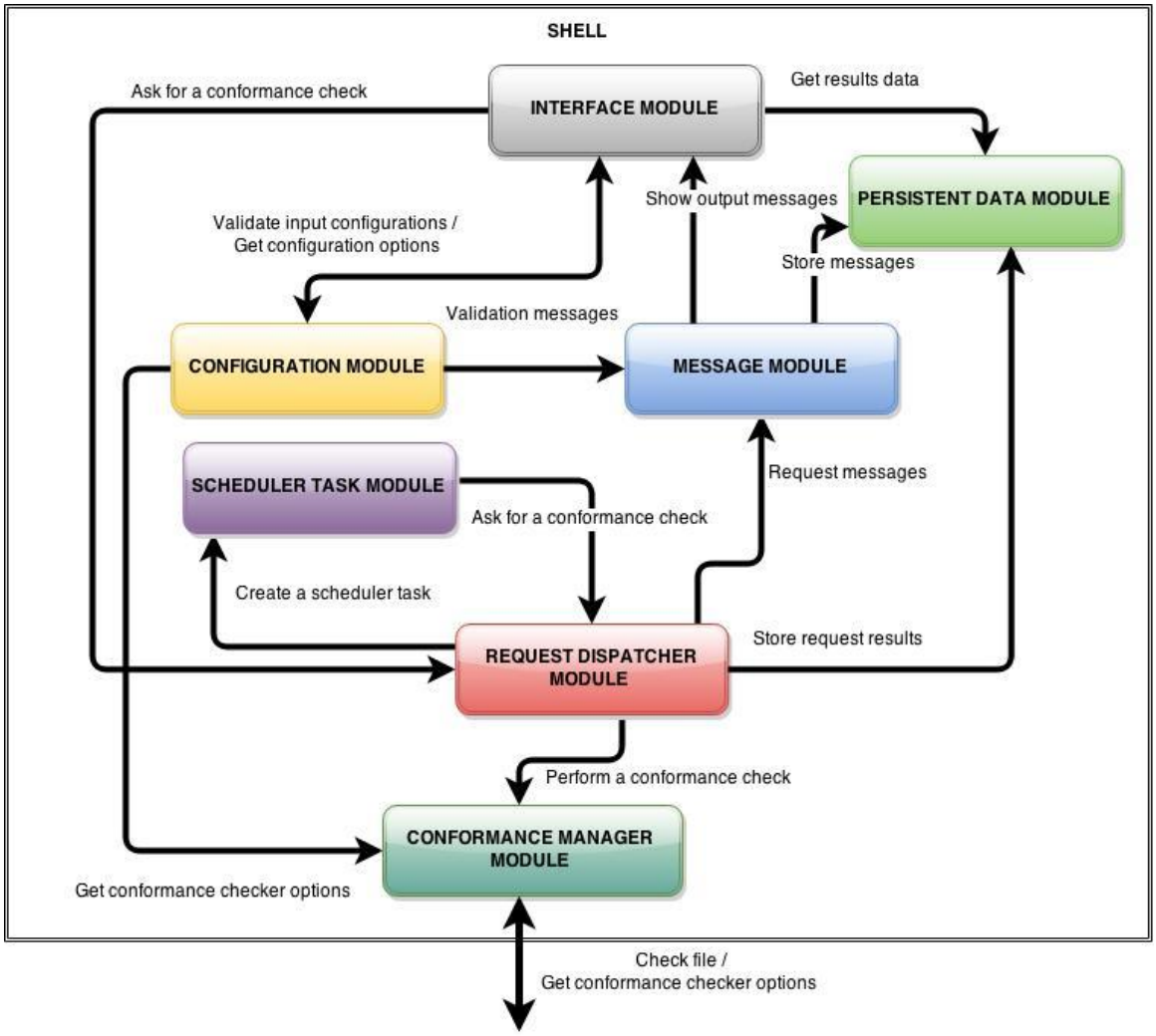
- **Commons CLI**: provides an API for parsing command line options passed to programs. It's also able to print help messages detailing the options available for a command line tool. This component is suitable for the command line interface.
- **Common Compress**: defines an API for working with ar, cpio, Unix dump, tar, zip, gzip, XZ, Pack200, bzip2, 7z, arj, lzma, snappy, DEFLATE and Z files. This component will be used to process compressed input sources.
- **Common Configuration**: provides a generic configuration interface which enables a Java application to read configuration data from a variety of sources. Will be used to set the general configuration parameters for the DPF manager.
- **Common Discovery**: The Discovery component is about discovering, or finding, implementations for pluggable interfaces. It provides facilities for instantiating classes in general, and for lifecycle management of singleton (factory) classes. This component discovered the new modules, objects added by developers to implement new functionalities**.**
- **Common Email**: Library for sending e-mail from Java. Used to inform about a scheduled conformance check action.
- **Common Launcher**: Cross platform Java application launcher.
- **Common logging**: Wrapper around a variety of logging API implementations used to log messages.
- **Common Validator**: Framework to define validators and validation rules in an xml file. Used to validate the XML input configuration and the XML report generate by the conformance checker.
- **Common VFS**: Virtual File System component for treating files, FTP, SMB, ZIP, WEBDAV and such like as a single logical file system. Used to support remote input file sources and remote store files.

All the Apache Common components are distributed over Apache commons 2.0 license that ensure the compatibility with DPF Manager licenses.

3.1.2. Shell 3rd party libraries

The Shell architecture is decoupled in modules where each module has a well-defined functionality.

In the following section we describe where in the shell architecture we going to use 3rd party libraries and the purpose of this libraries.

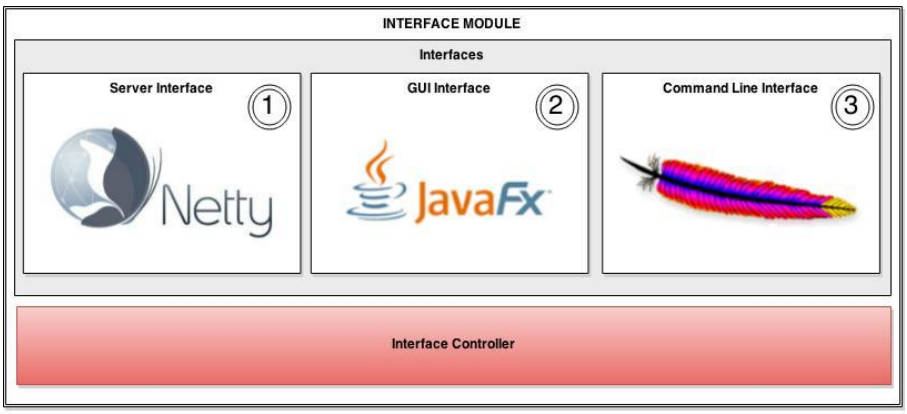


Picture 1 Shell modular architecture

The Interface Module

The Interface Module provides the access to the DPF manager functionalities. We understand that develop a server interface and a user graphical interface is completely out of PREFORMA project because the amount of time and money that could carry, as well as compromise the

stability of the hole project. In other to provide the interfaces required by PREFORMA we choose the following libraries.



Picture 2 Interface Module libraries

1. **Netty**. Netty is a NIO client server framework which enables quick and easy development of network applications such as protocol servers and clients. It greatly simplifies and streamlines network programming such as TCP and UDP socket server, and large file transfer. Netty is a mature framework, widely used (see: http://netty.io/wiki/adopters.html) with an extensive documentation.

   We also evaluate the use of other frameworks like Jetty. Jetty is more appropriate when you want to provides a Web server with a javax.servlet container to deploy WAR files while Netty is more suitable for a true Client-Server application.

   In this module Netty implements the server communication.

2. **Java FX**. Included in Java 7 JDK, JavaFX is a set of graphics and media packages that enable developers to design, create, test, debug, and deploy rich client applications that operate consistently across diverse platforms. With this framework we could create and attractive user interfaces for the standalone version.

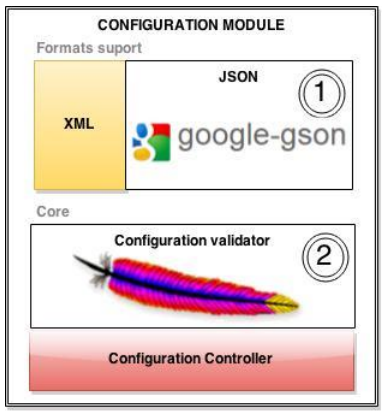
3. **Commons Cli**. Included in Apache Commons the Cli component provides the tools necessary for parsing command line options to be interpret.

Configuration Module

Java JDK 7 has the necessary tools to serialize XML files but not include tools to serialize JSON therefore we need a library to supply this lack.

1. **Google-gson**. One of the most used libraries to work with JSON objects in Java. Google-gson can be used to convert Java Objects into their JSON representation. It can also be

used to convert a JSON string to an equivalent Java object. Gson can work with arbitrary Java objects including pre-existing objects.

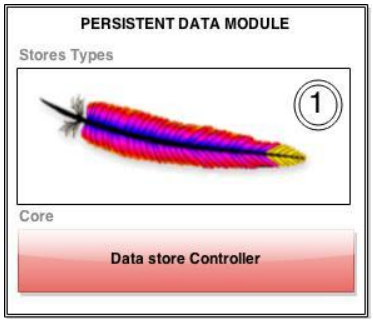


Picture 3 Configuration Module libraries

2. **Common validator**. Verifying the integrity of the input configuration files to ensure a successful file check is crucial. Although Java JDK7 provides a way to validate a XML through and XSD Schema, this library provides a set of utilities to validate different kind of data formats.

Persistent Data Module

1. **Common logging**. The most famous library for logging in Java. The use of this library will ensure that we are able to store the logging persistent data in our file system.



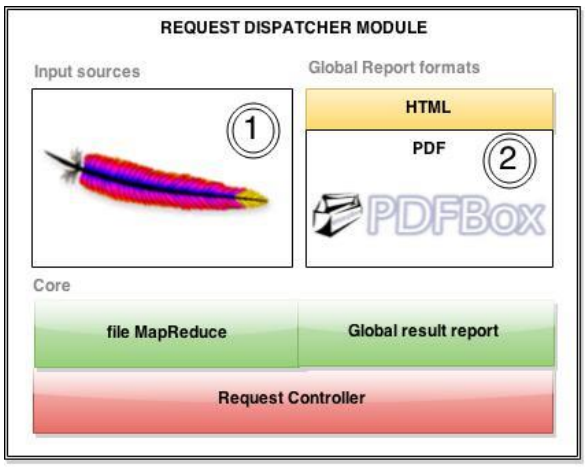
Picture 4 Persistent Data Module

Request Dispatcher module

This module is responsible to read the input source. We need to read as most input sources as possible to give the DPF manager flexibility to be suitable in different environments.

1. **Common VFS** and **Common Compress**. The common VFS give access to a large number of remote file systems and sources (see: http://commons.apache.org/proper/commons-vfs/filesystems.html) and with Common Compress we could process the most common compressed files formats.

2. **PDFBox**. Open source Java tool for working with PDF documents. This library allows creation of new PDF documents, manipulation of existing documents and the ability to extract content from documents. Used to generate a global report with templates.

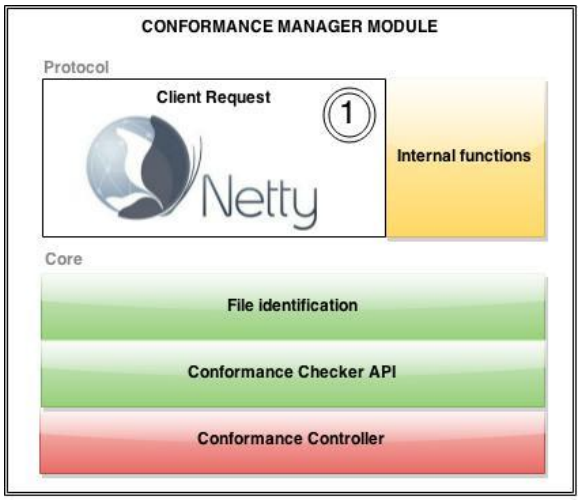


Picture 5 Request dispatcher module library

Conformance Manager Module

The conformance manager module is able to interact with the build-in Conformance Checkers as well as remote Conformance Checkers in a DPF manager server.

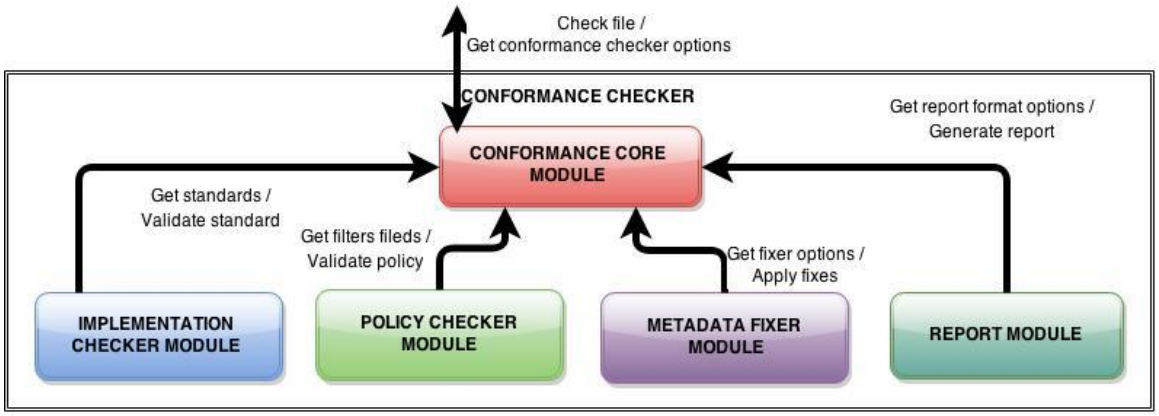
1. **Netty**. In this module Netty is used to implement the client communication with the server implementation.



Picture 6 Conformance manager module library

3.1.3. Conformance Checker without dependencies

The Conformance Checker internal architecture is structured in modules. All of them free of 3rd party libraries dependencies.



Picture 7 Conformance Checker modular architecture

In the document "EASY INNOVA – 3.Technical specifications.pdf, section 4, page 30 " the object structure of TIFF file format is described. This structure is used to store the TIFF content to be validated through the implementation checker and policy checker.

This section extends this content and describes how to handle with complex information inside the TIFF format.

Exif

The "ISO 12234-2 Electronic still - picture imaging - Removable memory Part 2:TIFF/EP image data format" define how to include Exif (Exchangeable image file format) information inside a tiff file. This information is structured inside an IFD structure and the content is described inside a new set of tags. Too handle with this information is extremely easy because we can only have to implement this new extension tags and the TIFF format reader will be able to read this information.

XMP

Out of the baseline 6.0 tags definition there is an extended tag to store XMP information.

Code	Name	Description	Source tag	Type
700	XMP	XML packet containing XMP metadata	Extended	BYTE

This information is stored inside the TIFF using and XML format, using the default XML Java JDK libraries we are able to read the content. There is no need to include the Java framework provided by Adobe to work with XMP content.

IPTC

The “ISO 12234-2 Electronic still - picture imaging - Removable memory Part 2: TIFF/EP image data format” define how to include IPTC (International Press Telecommunications Council-Newspaper Association of America) metadata inside a TIFF file using the following tag.

Code	Name	Description	Source tag	Type
33723	ITCP/NAA	IPTC-NAA metadata.	TIFF/EP	UNDEFINED

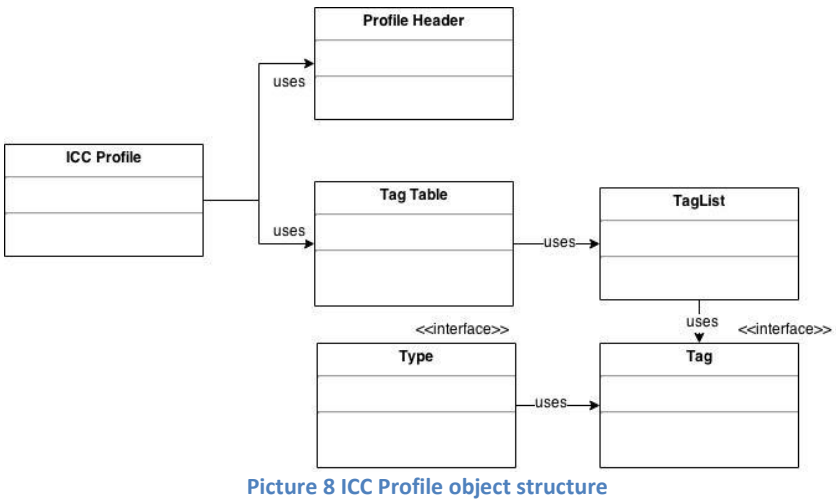
The information inside the ITCP tag is encoded using an XML format so the Java Libraries included in de JDK are enough to read the content.

ICC Profile

The ISO “ISO 12234-2 Electronic still - picture imaging - Removable memory Part 2: TIFF/EP image data format” and the “ISO 12639:2004 Graphic technology -- Prepress digital data exchange -- Tag image file format for image technology (TIFF/IT)” define a tag to embedded ICC colour profile.

Code	Name	Description	Source tag	Type
34675	ColorProfile	ICC profile data.	TIFF/EP and TIFF/IT	UNDEFINED

The ICC Colour Profile follows a well-defined structure described in the “ICC.1:2010 Specifications Image technology colour management – Architecture, profile format, and data structure“ version 4.3.



Picture 8 ICC Profile object structure

Using the ICC profile object structure we will be able to work with ICC profiles embedded inside the TIFF format.

The most relevant part of the ICC Profile is the profile header where we could find the general information about the profile; profile class, colour space, signature, Profile ID…. This information can be filter in the policy checker using the corresponding selector implementation as it is explain in the document “EASY INNOVA – 3.Technical specifications.pdf, section 4, page 33 ” .

All the information collected in these tags is send to the Metadata Fixer Module. The Metadata Fixer module verify the integrity of this data and make auto fixes over it if there are errors. Then apply the metadata changes required in the XML configuration file as it was explained in “EASY INNOVA – 3.Technical specifications.pdf, section 4, page 34 ”.

The reporter module includes all the metadata collected and fixed inside de XML report structure. All this metadata could be used to fill metadata standards structures like PREMIS or METS.