Project Octopus

an open community curated provenance system to track and encourage the cultural flows of creative works online

H2020-ICT-19-2015

Topic: Technologies for creative industries, social media and convergence

Action: Research and Innovation

Participant No.	Participant organisation name	Short	Country
1 (Coordinator)	Kennisland	KL	NL
2	Peer Practice	PP	DE
3	Commons Machinery	CM	SE
4	iRights.Lab	IR	DE
5	Instituut voor Informatierecht	IViR	NL
6	Klokan Tech	KT	CH
7	Tribe of Noise	ToN	NL
8	National Library of Luxembourg	BnL	LU
9	Wikimedia Deutschland	WMDE	DE

Table of Contents

1. Excellence

The Internet is a media-sharing technology. Images, sounds and video are responsible for the bulk of the traffic on the Internet. This enables successful business models like on-demand video, audio streaming and stock photo sites. These business models are Intellectual Property Rights (IPR) intensive industries which contribute 3.2% of employment and 4.2% of GDP in the EU.[1] These industries, including the creative industry and the publishing industry, rely on information about the source, owner, copyright status and permissions of reuse of creative works. This information – provenance information – is often missing. Provenance information is an essential part of the creative process, both on and offline, foremost because of the interconnectedness of culture.

Culture does not thrive in isolation, it builds on top of what already exists. The digital world offers a multitude of opportunities to remix cultural works, a lot of creative products have been developed for that digital world like fast media-sharing platforms such as Facebook and Tumblr. This generates extremely powerful creative incentives, in which unprecedented amounts of users take part. It can, however, prove difficult to legally build upon these new works when provenance information is lacking. This problem hinders (commercial) co-creative processes as creators do not have access to reliable information about which works can be reused under what permissions.

Problem: Tools to share and publish media on the Internet have been developed extensively, however there are no platforms to track in what way media is shared across different online publishing environments. It is very difficult to find provenance information of media, especially when it is shared outside of its original publishing context.

Project Octopus develops a cultural web observatory that enables users of the Internet to trace different publishing contexts of a work. It is a community-based project that engenders opportunities for creators of digital cultural works to create new creative works, improve access and reusability of digital cultural works, make content retrieval easier and entice new possibilities for interaction with creative content.

At the moment, the largest bulk of media is shared on a few large web silos, or walled gardens, such as YouTube, Facebook and Instagram.[2] Ownership and permission over contributed media is settled by use of license agreements. On these platforms information about the

[1] According to 'Intellectual property rights intensive industries: contribution to economic performance and employment in the European Union' by OHIM and the European Patent office: http://ec.europa.eu/internal_market/intellectual-property/docs/joint-report-epo-ohim-final-version_en.pdf.

[2] Youtube adds 300 hours of video every minute (https://www.youtube.com/yt/press/statistics.html), Instagram has over 300 million active users (http://blog.instagram.com/post/104847837897/141210-300million), and Facebook has 890 million active users each month (https://newsroom.fb.com/company-info/).

creator is often separated from the media file, either because the users did not add it during upload or because the platform does not provide options to add provenance information. Moreover, when media is shared outside of these platforms, provenance information often gets lost. Consequently, at the moment only a fragmented story of the use of creative works on the Internet is available. A photo can first be published on a smartphone, then moved to Tumblr, then to Twitter and Flickr and might then be republished, copied, edited and reused on multiple personal blogs. No one has thus far attempted to systematically put together a picture of the flow of cultural works across the web.

Project Octopus develops a first ever cultural web observatory of cultural flows. The web observatory indexes the use of media on the web. This system gathers and links flows of cultural data online. It allows users to answer questions about that data to gain deeper insight into the way media is used and shared, and to come to understand its reach and adoption rate on the web. Also, provenance information is vital for legal use of creative works. The creative and publishing industries, essential sectors in our economy, need tools to be able to follow IPR legislation. Project Octopus tracks where media is used on the Internet and provides provenance information by:

- ➔ Collecting use and provenance information.
- ➔ Creating a dashboard for creators dedicated to tracking use of works.
- ➔ Sourcing provenance information from across the web.
- ➔ Being a de facto copyright registry.
- ➔ Enabling machine-to-machine interactions (APIs).
- ➔ Providing a mechanism to enrich, correct and search data.
- ➔ Enabling community enhancement and interactions.

Create, Access, Retrieve and Interact

Project Octopus is a **community-based** project that generates opportunities for creators on the web to create new creative works, improves access and reusability of digital cultural works, makes it easier to retrieve content and entices new possibilities for interaction with creative content.

Creation of new content drives culture; content helps to understand and reflect upon society. Co-creation and building on existing culture makes society thrive. Our copyright framework[3] is the product of a policy that protects creators and stimulates the creative sector. Because of exclusive copyrights on a creative work, rights owners can decide where their works can be shared and who can copy or build upon their works. Correct provenance information is crucial to exercise and respect copyright as it allows use without fear of infringing third-party rights. On the Internet there is a lack of this information. Consequently the creative industry is hindered in co-creative processes as it does not have access to reliable information about which works can be reused under what permissions. Simultaneously we see that publishing practices differ from the principles that underlie the copyright framework. Project Octopus provides insights both for creators who want to exercise their rights, as well as for online

[3] Part of the EU aqcuis, see http://ec.europa.eu/internal_market/copyright/acquis/index_en.htm.

publishers to have insights to stimulate co-creative processes. Octopus thereby presents a unique perspective on the publishing realities of the web.

Access to content is for users often not enough. Users, like online publishers, want to be able to remix content and build upon found digital material. Often reuse permissions are given to single publishing platforms or provided by means of licenses. Project Octopus not only indexes media on the Internet, it provides tools to determine the rights status of a work, it facilitates easy identification of rights owners and stores licensing information when known. **Content retrieval** of works that have been transformed or build upon is not easy on the web. Project Octopus provides provenance information with matching media files based on media recognition software to assist online publishers to explore remixed works.

Interaction with content takes place in different ways; either via content platforms such as Flickr, Wikimedia Commons, social media sites like Twitter or Instagram, or via web browsing. A network that shows where creative works are published online creates a new unique way of interacting with media. Media can be explored by visiting the multitude of locations where it has been published in the past. This leads to unique new ways of traversing context where media is published by creating links between previously unconnected contexts.[4]

Value of culture online is hard to determine. Lack of provenance information leads to inadequate and inaccurate filtering, crediting, licensing, discovery, and use of creative works. As such, the value of media on the Internet cannot be taken advantage of to its full potential. Access to reliable provenance information helps to unlock the potential of the wealth of culture that is shared online (both in and outside of the EU).

Communities

Online publishers, both professional and hobbyists, encounter problems caused by inadequate provenance systems. As problems differ per community, these publishers are categorised into 3 communities:

- ➔ Creative communities.
- ➔ Galleries, libraries, archives and museums (GLAMs).
- ➔ Remix communities.

Creative communities

Creative reusers on the web (both individuals and organisations working in the creative industries) publish content and build on top of content that is available online. To stay within legal boundaries, materials can only be reused if they are in the public domain, licensed under an open license that gives permission for reuse, or with explicit consent of the rights holder.

[4] Files on the Internet are mainly connected through hyperlinks. If there is no explicit link between two web pages there is very little chance that a user derives the link between one context and another. An open provenance repository with media recognition technologies draws new links between these contexts by use of the embedded media files. For example, a copy of the same image can be used on Wikipedia, a news article and personal blog. Traversing these sites through the shared media files provides a unique experience that differs fundamentally from the current user experience.

Provenance information is a necessary component of reuse for creative communities. Third parties that reuse their creative materials also have to ask for permission. It is important to have provenance information attached to this material that links back to them.

A lot of reused materials are found on media platforms such as Flickr, Tumblr, Vine or YouTube. It is not the prime interest of these platforms to provide users with meaningful provenance information that is useable outside the platforms.[5] As such, a handful of dominant web silos have created an environment in which society's creative works are underutilised and are shared without proper provenance information. Provenance information is a key component to maximise a work's value for its creators.

Furthermore, because copying and republishing of media is commonplace on the Internet, creative communities and publishers on the web need widely available technological methods to be able to track creative works as they move across the web. To do so it is necessary to be able to structurally and efficiently compare works. Currently there are no widely available technological methods for this.[6]

GLAMs

GLAMs that share (parts of) their digitised collection online have little structural insight into where their works are shared on the Internet. Some institutions do not publish their collections online because of this. Experience has shown that cultural heritage can reach to over 17.000 fold more audience on other platforms than their own sites.[7]

A lot of interaction with collections takes place on third party platforms, where embedding and copying is widespread. Tools for measuring use and providing correct attribution are often missing. As it is currently not possible to efficiently trace different publishing contexts, GLAMs struggle to connect to their communities and to measure the impact of digitised collections that are published online. Cultural institutions undervalue the impact of their own collections as they cannot determine their reach on the web.

Remix communities

Remixers of creative works likewise find themselves restricted by the lack of provenance information. Wikimedia projects, for example, cannot use much of our cultural heritage because of lacking provenance information. Especially when rights and reuse permissions of creative works are unknown. This evidently leads to less availability of – and access to – knowledge and creative content online.

[5] These platforms are optimised for great user experience within their platforms, and not for copying or interconnectedness with other platforms.

[6] Options that do exist like Google reverse image search and TinEYE rely on proprietary algorithms and software that do not connect search results to provenance information. In addition, the provenance information accumulated by these tools is not available for other purposes.

[7] As shown in a Dutch impact assesment on the reach of its national archive on Wikipedia in 2012: https://commons.wikimedia.org/w/index.php?title=File:BvdT_effectmeting_2012.pdf&page=4

For these communities orphan works and copyfraud are large problems. Both of these are results of insufficient reliable provenance information. (Involuntary) copyfraud – no attribution at all, or wrongful claims to copyright – is abundant on the Internet, as it is difficult for users to reliably retrieve information on the copyright owner of an item. Orphan works lack provenance information altogether. This information might however be present elsewhere on the Internet. In the current situation, it is very costly for users to collect information about a work, if possible at all.[8] Orphan works form a problem to the extent that the EU has written special directives on mitigating these issues.

Project Octopus

Project Octopus combines provenance information of works from GLAMs, the web and the expertise of a community to mitigate current problems, to reduce the timely and costly exercise of retrieving provenance information.

In sum, culture cannot thrive to its full potential in the digital environment. Combining latent provenance information from across the web – information from web silos, GLAMS, and individual websites – is a first step to help Internet publishers to determine whether they can reuse material. While not part of this project the open infrastructure of Project Octopus allows other parties like collective management organisations (CMOs) to more efficiently execute their missions. This web observatory of media empowers the creative industries to more easily build upon works by others and the media industries to track their publications.

1.1 Objectives

Project Octopus creates an operational open repository (1) of community curated information in a provenance system (2). The project does scholarly research to gain a better understanding of the relationship between provenance information and the copyright framework (3). It enables the use of creative works by establishing a whitelist of both openly licensed information as well as public domain materials (4). It develops open source technologies to be able to identify works (5). This leads to the first ever web observatory of cultural flows (6).

Objective 1: CREATE a community curated provenance system

Project Octopus provides new ways to discover culture and understand and visualise cultural flows. After mass ingestions, its database is opened up for curation by a community, which results in the availability of reliable documentation of the provenance of works, relationships among works, and uses of works on the Internet. It gives insight into where works are used, who their rights holders are and whether a work can be (freely) reused by third parties.

[8] See http://europa.eu/rapid/press-release_MEMO-12-743_en.htm?locale=en

The project indexes the needs of creative communities, remix communities, and GLAMs. It develops tools and services to serve these communities, and creates a strong active community that furthers the project's work.

- ➔ Create an open source provenance system.
- ➔ Develop community tools and value adding services.
- ➔ Engage with several communities to connect these systems and services with their prospected users.

This objective is mainly accomplished within the tasks of the development work packages WP1 and WP2, and WP4 (Community engagement) as well as in close collaboration with project OutOfCopyright.eu.[9]

Objective 2: ENGAGE communities to curate and add provenance information

Communities are crucial in data curation projects. Communities like Tribe of Noise, Wikipedia, Wikimedia Commons, WikiData, and MusicBrainz rely on their communities to curate and increase the quality of their data. Active participation of user groups is a requirement for a successful project.

The project engages three identified user groups: creative communities, remix communities and GLAMs. It gathers the needs of these communities for provenance information. It reaches out to these user groups and introduces the developed tools, ensuring wide dissemination and promoting engagement.

This objective is mainly accomplished based on the work in WPs 1 and 2 on development with input from WP 4. After a stable release of their input WP 4 reaches out to existing communities and starts developing a community that curates provenance information.

Objective 3: RESEARCH the position of provenance information in the current copyright system

The relationship between provenance information, copyright formalities, other types of rights management information and existing private enforcement systems – such as Content ID[10] – is unclear. It is the project's objective to research in what way an open online provenance repository can improve rights clearance operations, and how it can improve existing open

[9] OutOfCopyright.eu presents research and tools developed on the rights status of works within Europe. It helps to determine whether a work is still restricted by intellectual property rights given an EU jurisdiction and provides insight whether new intellectual property is created when digitising analogue works.

[10] Content ID is a proprietary system developed by YouTube to match audio and video files based on similarities to new uploads of the platform. It helps to scan for copyright infringements by YouTube's users and offers options to rights owners when a match is found. For more information go to https://youtube.com/t/contentid.

licensing systems that currently suffer from weak links between the licenses, licensors and the licensed works.

A related objective is to research the requirements for use of a provenance system in the context of current copyright framework. The project studies the legal barriers of such systems in relation to international agreements, proprietary rights managements systems, and the legal consequences of claiming ownership of a work in such systems.

- ➔ Research personal data protection issues related to open online repositories for provenance information.
- ➔ Research provenance information as a driver for more efficient licensing and rights clearance.
- ➔ Research the legal validity of provenance information in the context of the copyright framework.

This objective is accomplished in WP3 and relies on the expertise of the IViR, KL, PP and IR. It builds upon their previous work as affiliates of Creative Commons, OutOfCopyright.eu, copyright reform advocacy and scholarly research.

Objective 4: ENABLE reuse by establishing a whitelist of works in the public domain and reuseable creative works

Project Octopus improves the availability of reusable works by enabling (automatic) rights clearance operations as it indexes openly licensed information. Its objective is to create these possibilities within by developing whitelists of public domain works and providing the ability to store structural licensing information.

At the moment there is no structural store for information about licenses and reuse of works that exists outside of specialised media platforms. As media is copied and shared on the web this information often gets lost. There is usually no way to trace the publishing history back to their original platforms. Creating and generating lists of freely reusable information increases the value of the easily reusable works.

- ➔ Create a whitelist via Project Octopus.

This objective is accomplished in WP2 and is based on research and development done by the IViR, KL and The National Library of Luxembourg. It makes use of the platform OutOfCopyright.eu that has been previously developed by these three partners.

Objective 5: IMPROVE openly available implementations of media recognition algorithms

There are plenty of services and algorithms available to determine if two media files are the exactly same or similar, and to what extent the two show similarity. Most of these services rely on proprietary algorithms and software. Examples of these are Google reverse image search and TinEYE. Closed source products and proprietary algorithms lead to dominant market positions where these services control what information is shared about creative works. It does not enable collaborative approaches to determine where works are used on the Internet. To empower online publishers, it is the objective of the project to create openly available implementations of media recognition algorithms.

Online publishers can apply this media recognition software in their own endeavours, creating new business opportunities, and allows them to further the general objectives of this project. The project researches openly available algorithms for media recognition of images, audio, and video. It implements these algorithms as open source, and keeps in mind that a general public should be able to make use of what is developed.

➔ Implement media recognition algorithms as open source.

This objective is accomplished in development work package WP2 in close collaboration with WP1 and relies on the past experience of our partners KT and CM, the latter of which is the founder of an open source image recognition service Elog.io.

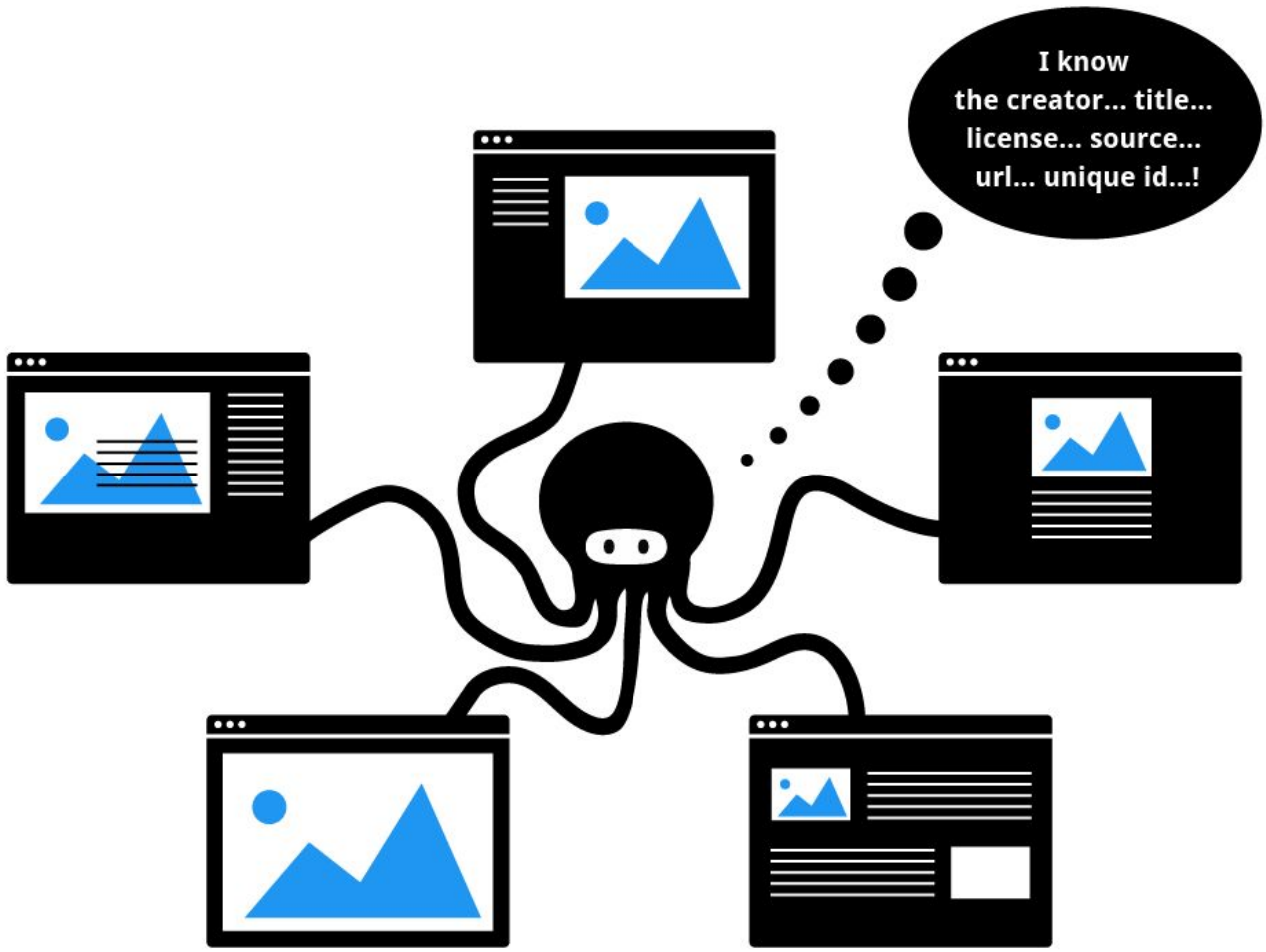
Objective 6: PROVIDE visibility into how creative works are reused across the web

Project Octopus maps movements of creative works on the web. Currently there are no tools that collaboratively index which creative works can be found on the Internet. The community-based provenance system of Project Octopus makes it possible to visualise the online different locations where works can be found.

This provides new insights into the co-creative processes of sharing, (re)publishing and adopting vast amounts of works already available online. These insights provide value for all the project's user groups. For GLAMs it becomes insightful in what way their collections are used on the Internet, it helps them identify new audiences and helps them make informed policy decisions. Creative communities benefit by seeing how their works are being reused and identify their fans. Remix communities can traverse the movements of creative works to find similar resources for their creative products and find new contexts of reuse.

➔ Develop a dashboard to visualise the use of works on the Internet.

This objective is accomplished in the overall project. It relies on the technological development of WP1, the additional services of WP2 and our community outreach in WP4.



Octopus is a cultural web observatory that enables users of the Internet to trace different publishing contexts of a creative work by collecting latent provenance information.[11]

1.2 Relation to the work programme

The Octopus project directly addresses the **research and innovation action** of the Horizon 2020 19/2015 call with the specific topic **technologies for creative industries, social media and convergence.**

From ICT 19/2015 – Technologies for creative industries, social media and convergence

[11] Project Octopus' logo 'Octopus' is created by Jason Grube and licensed under a Creative Commons Attribution license. Go to https://creativecommons.org/licenses/by/3.0/us/ for the full license information.

Thanks to ubiquitous technology adoption, widespread use of mobile devices, broadband Internet penetration and increasing computing power the consumption of content anywhere, anytime and on any device is becoming a reality. Consequently, **developments related to content creation, access, retrieval and interaction offer a number of opportunities and challenges, also for the creative and media industries.**

The ease of content creation in the digital environment and digital publication provides us with novel problems concerning the reuse and reach of these works. Most of the publication technologies and publishing platforms focus on content creation and not on the underlying provenance information of that content. Hosting platforms are simply not primarily concerned with providing correct and extensive provenance information of content after it has been published.

This leads to inadequate and inaccurate filtering, crediting, licensing, discovery, and use of works after their initial publication and on other platforms. It hinders the creation of derivative works, proper access, discovery and retrieval of creative works on the Internet. The project directly addresses these challenges by separating the publication platforms from their provenance information. Project Octopus collects reliable provenance information both to enable and to track reuse. The project allows provenance information to be retrieved independant from publication platforms.

In order to keep pace with the trends and remain competitive, **those industries need to explore new ways of creating and accessing content**.

Creative industries and media industries are responsible for a large section of Europe's jobs.[12] To remain competitive and to optimise the use of our creative products we need to be able to find and track the works that we publish on the Internet. Furthermore, the ability to trace back who is the rights holder enables proper use of creative works. The first step to achieve this, is by creating a database with correct and extensive provenance information.

It is extremely complicated to automatically recognise a media file as similar or the same as another media file. Private parties like Facebook, Google, etc. have successfully implemented algorithms to identify similarities between media files. However, these closed source proprietary implementations do not serve the smaller market participants. This project makes new technologies freely available as open source for the entire creative and media industries to be able to compare media files.

[12] See http://documents.epo.org/projects/babylon/eponet.nsf/0/8E1E34349D4546C3C1257BF300343D8B/$File/ip_intensive_industries_en.pdf

From ICT 19/2015 – Technologies for creative industries, social media and convergence – Research and Innovation

> Research in **new technologies and tools to support creative industries** in the creative process from idea conception to production. The proposed tools should **explore the potential of technology to enhance the human creative process** from the expression of ideas to experiment solutions.

The Internet provides a frontier between public/private & commercial/non-commercial activities. It is important to have provenance information to prevent unnecessary lawsuits. Many organisations in the creative industry typically work online and make use of materials that have previously been digitally published. They build on top of culture that circulates both on and offline. Currently, the creative industries are prone to breaking legal rules in the process of reusing materials online. If they want to stick to legal boundaries, they have to invest a lot of time and energy into tracing provenance information to check whether a work can be reused. This takes their time, energy and resources away from their creative process.

Project Octopus develops an open centralised platform that reduces transaction costs, which helps all market participants. Project Octopus minimises barriers for reuse, so that creators can focus on their process instead of the legal frameworks they need to operate in. The larger aim of the project is to encourage and support a culture of (co-)creation and remixing.

> **Collaboration and user-community interaction** should be improved **based on research** leading to a deeper understanding of the dynamics of co-creative processes.

In the initial stage of the project, research is performed to map the needs of several user communities. Three different user communities have been identified: GLAMs, creative communities and remix communities (such as the Wikimedia community). The problems they encounter and their desires and needs are identified and taken into account when building the repository.

The overall product is an observatory of creative works. This provides unique insights into how communities publish and reuse creative material on the Internet. A web observatory generates a deeper understanding of how creative works travel on the Internet. It is both an observatory for fans to find similar works, creators to track their works and policymakers to make informed policy decisions.

> The tools should be **cost effective, intuitive, and be demonstrated in real-life** environments **relevant** for the creative industries.

Project Octopus shares all of its products as open source, open standards and open data. This encourages collaboration and ensures that product results remain reusable by third parties. This makes building upon Project Octopus effective, and allows for further community contributions.

Structurally creating and using open data and publishing all its products as open source increases transparency and efficiency of the project both in the short and long term. Openly licensed products ensure that researched methodologies can be used by all, that developed software can be embedded in other projects and that other parties have the opportunity to build their own services on top of the infrastructure and services offered by Project Octopus.

The goal of Octopus is to enhance opportunities for creative industries and other stakeholders in the digital environment by providing the opportunity to collect and trace provenance information of works online and by creating clarity about provenance of digital material, which makes reuse online without infringing intellectual property rights easier and less time-consuming.

Overall programme – "Content technologies and information management"

> Provide professionals and citizens with new tools to model, analyse, and visualise vast amounts of data from which to **extract more value**, to make an intelligent use of **data coming from different sources** and to **create, access, exploit, and re-use all forms of digital content** in any language and with any device**.**

Knowing where creative works are used and understanding the general flow of creative works can transform many business practices, both for creatives and media industries. The developed system becomes a comprehensive and reliable source of provenance and usage information for digital works, enabling new opportunities to establish metrics and discovery in the digital environment.

In its current form the Internet provides a very efficient platform for content distribution (moving content around) both within specific platforms (such as YouTube, Facebook or Wikipedia) but also between these platforms and the rest of the Internet. As long as content remains on a specific platform the proprietors of these platforms are able to track the movement of content and to keep track of the associated provenance information. This becomes much more difficult as soon as content moves across platforms and the rest of the Internet. Project Octopus proposes to build an open repository of provenance information that allows this information to travel together with media objects as they move across the Internet.

The project enables creative makers to register creative works, allows users to look up provenance information based on a file found at an arbitrary location on the web, and promotes and protects the public domain and publishers of open content by creating a whitelist of freely reusable works. These gain value, feasibility and commercial opportunities

(for third parties) as the developed system scales. Finding reuse and provenance information on material on the Internet is essential to the digital creative process.

Open governance, community curation and experienced community outreach based on research helps to determine the use of creative works on the Internet. It provides a necessary structural link between common practices for Internet publishing and the creative processes based on media found on the Internet.

1.3 Concept and approach

> ***"When the time is ripe for certain things, they appear at different places in the manner of violets coming to light in early spring."*** [13]

Provenance systems have been developed since the early 2000s. These previous attempts to build a functioning and widely used provenance system have proven to be unsuccessful. In 2014 multiple partners in the project realised that the technology and the need for these systems have advanced sufficiently to warrant a restructuring of the problem, given today's challenges in our digital society, and to collect those parties that are working on these issues. After thorough research, which resulted in a paper[14], Project Octopus identified demands and approaches to a provenance system. It now gathers efforts and has built a strong and capable consortium to tackle this problem from a novel perspective.

[13] Farkas Bolyai to his son Janos, urging him to claim the invention of non-Euclidean geometry without delay. See https://en.wikipedia.org/wiki/Multiple_discovery#Quotations

[14] The paper is available at http://about.project-octopus.org/2015/01/30/hello-draft-background/

1.3.1 Approach and methodologies

Project Octopus' general approaches are grounded in thorough research into previous attempts for provenance systems online.[15] Its approach and methodology serve the overall aim to support processes of the creative industries and to assist the media and publishing industries to effectively work with digital media.

Enable others in their creative processes

Digital creative processes can be restricted by legal, social and technical infrastructures. Barriers that arise because of a lack of provenance information are a burden for digital creators. Project Octopus is designed to remove barriers by providing the opportunity to track works online.

Project Octopus is inspired by the Creative Commons (CC) license infrastructure. CC is the largest infrastructure provider of open licenses for creative content. CC licenses are a simple, standardised way to grant copyright permissions to creative works. It formalises permission of reuse that help creative professionals to allow others to build upon their work and find works to build upon.

Innovate the creative and publishing sectors by making all (provenance) databases and products openly available

Project Octopus creates and works with Free Libre and Open Source Software (FLOSS) to ensure the usefulness of the project's outcomes. Likewise, the project publishes its repository as open data. This allows others to build upon the developed products. It contributes to the general innovation of other market parties in the same field. Especially Octopus' media recognition development is of great benefit for work in neighbouring issues and initiatives.

Our data model of provenance information as well as our implementation of recognition algorithms are published as open standards. This data model can then be referred to, embedded in other projects and publicly commented upon. When derivative implementation of algorithms is made within the project we ensure that these are published as open standards as well.

The entire database of provenance information that Project Octopus collects is published as open data, for others to reuse and build upon. This allows other interested market parties to work with this dataset to achieve goals that are not in this project or to improve the existing information and implementation.

[15] Resulted in a paper on provenance systems. This paper shows that all previous attempts to build a functioning and widely used provenance system have proven to be unsuccessful. It indicates that ongoing efforts meet the same fate. However, it is useful to develop a new provenance system, which is dedicated to tracking use of works and sourcing provenance information from across the web: that allows everyone to identify and track digital content and related licensing information while avoiding building another locked-in database. See http://about.project-octopus.org/2015/01/30/hello-draft-background/

Empower communities of creators and users

Our open approach leads to the most important aspect of the project: to support the co-creative processes of creative professionals on the Internet. It does so by actively engaging with communities who work on provenance information to improve discovery, proper use of licensing, reduction of copyfraud, etc. And also by providing a wider community with the tools to build upon this core foundation or to create own similar projects based on this work.

Grounded in research into provenance in the copyright framework

Provenance information has the potential to improve the way copyright is exercised in the online environment. Project Octopus aims to make a step towards realisation of this potential.

Based on extensive research, a benchmark of criteria is identified according to which collected provenance information can be deemed reliable enough to be used in licensing and rights clearance contexts. Research determines the specific criteria for provenance in the context of copyright (whitelist, improving open content licenses, conflict resolution).

Additional research explores the relationship between provenance information and copyright formalities, other types of rights management information and existing private enforcement systems (such as contentiD).

1.3.2 Linked research and innovation activities

All consortium members are established innovators in the digital (heritage) field. Members have worked and still work on related research and innovations, with various Technology Readiness Levels (TRL), that are fed into Project Octopus.

The project's core infrastructure and user interfaces are based on proven (TRL 9) web applications and web services technologies, including HTML/CSS/Javascript for the user interfaces, JSON and REST for the services, and Apache/Ruby/PostgreSQL on the backend. The project aims to have an overall technology readiness state of at least TRL 7.

Provenance systems

Project Octopus builds on an explorative study performed by PP and KL. The study resulted in detailed analysis of provenance systems.[16] The study analyses all previous attempts to build a functioning and widely used provenance system and evaluate their successes and failures. The main point of failure is locked-in databases and proprietary software. The study concludes that an open online repository for provenance information that does not have copyright registration as its core goal is desired and viable. Based on this study a proof of concept was developed (TRL 3).[17]

The project builds on previous research undertaken by Commons Machinery as part of the Elog.io service. Elog.io developed an infrastructure for conveying provenance information of

[16] See http://about.project-octopus.org/2015/01/30/hello-draft-background/

[17] See https://project-octopus.org/

any type of digital work, and demonstrated how this could be used with plug-ins within a browser environment. Research on data models and image comparison that was undertaken to build Elog.io are fed into this project (TRL 6).

Finally the project builds on expertise developed by KT in comparing large sets of media files in order to recognise similar or the same analog work that has been digitised by different cultural heritage institutions (TRL 4).

Europeana and rights infrastructures

KL, IViR and the National Library of Luxembourg collaborated in Europeana Awareness[18] on OutOfCopyright.eu. The platform is the product of research and tools developed over the course of five years on the right status of creative works within Europe. It provides Public Domain Calculators and maps on rights after digitisation to help answer the question whether a work is in the public domain or not. Its research on rights determination and calculators results are fed into this project (TRL 7).

The same partners have worked in Europeana Connect on the Europeana Licensing Framework. The Europeana Licensing Framework standardises and harmonises rights within large collections of cultural heritage. Comprised of four elements, the Licensing Framework aims to bring clarity to a complex area, and make transparent the relationship between end users and cultural institutions (TRL 9).

Creative Commons

Key personnel of participants (KL, PP, CM, IR, IViR, and BnL) are, or have been, involved with the Creative Commons (CC) movement[19]. CC provides a simple, standardised way to give the public permission to share and use creative works – on conditions of their choice. CC licenses let makers easily change their copyright terms from the default of 'all rights reserved' to 'some rights reserved'. Project Octopus adopts this approach and framework. (TRL 9).

WMDE and ToN rely heavily on Creative Commons licenses for their organisations. WMDE relies on them to support their communities in working on Wikipedia and other Wikimedia projects (TRL 9). ToN uses CC licensed music in their business model (TRL 9).

[18] Europeana Awareness was a Best Practice Network to publicise Europeana at a political level, promote its use by the general public, develop new partnerships and further encourage cultural institutions to provide content. See http://pro.europeana.eu/web/europeana-awareness for more information.

[19] Paul Keller (KL) is Public Lead for CC Netherlands and member of the Board of directors at Creative Commons. Catharina Maracke (PP) was director of Creative Commons International. Mike Linksvayer (PP) was Vice President and CTO of Creative Commons. Jon Phillips (PP) was lead developer at Creative Commons. Jonas Öberg (CM) was Regional Coordinator at Creative Commons for Europe. Lucie Guibault (IViR) is legal lead of CC Netherlands. Patrick Peiffer (BnL) is Public Lead of CC Luxembourg. Johan Weitzman (IR) is currently Regional Coordinator at Creative Commons for Europe.

1.4 Ambition

Project Octopus is the first independent attempt to create a web observatory of the flows of cultural material on the web (1) to enable heritage institutions to determine the value of their digital reach (2). Project Octopus can revolutionise policy making by providing tools to make informed and rationalised decisions based on the true value of media (3). Project Octopus believes that a trustworthy way of creating this observatory is to develop a community of provenance information enthusiasts (4). Finally, in doing so, the project innovates governance models of provenance systems (5).

Ambition 1: Create a web observatory of the flows of culture

Previous provenance repositories either functioned as copyright registries, or did not track the locations where media is used. The success of the project relies on the large scale of the database. The choice for a user perspective rather than a policy perspective is grounded in a thorough research on previous attempts for a provenance system. Engaging an active community encourages a different way of relating to the information in the system. The reliability of the provenance information is curated by a community of enthusiasts and stakeholders that track individual records, while GLAMs contribute their structured data on cultural works.

The observatory requires a state-of-the-art combination of technologies. Project Octopus approaches the relationship between copyright and publishing realities from an interdisciplinary perspective to get a full picture of cultural flows.

Ambition 2: Enable heritage institutions to determine their undervalued digital reach

Heritage institutions measure the reach of their collection based on visitor counts at their physical location and often also on their own online platforms. This is no longer the publishing reality of the web. Cultural heritage is aggregated by large projects like Europeana[20]. These aggregators allow others to use that data via APIs and distribution platforms. The use of the referenced media after it has been copied from those aggregators is not structurally tracked.

Heritage institutions also publish public domain works[21]. Public domains works are no longer restricted by copyrights or neighbouring rights. This means that permission is no longer

[20] Europeana.eu is an aggregator of cultural heritage data. It collects datasets from over 4.000 cultural institutions in Europe. In April 2015 information of more than 40 million records is aggregated.
[21] Europeana.eu has information of over 9 million public domain works in its database.

required for them to be used on the web. This leads to a wide dissemination of these creative works. This widespread is not tracked either and therefore not properly valued.

Not counting this dissemination means to undervalue the collections owned by our cultural heritage institutions. Project Octopus provides the ability for these institutions to track their collection online. This revolutionises the way these institutions publish and deal with their online presence. It makes more cultural heritage available online for creative use, but also helps researchers and educators in their efforts.

Ambition 3: Rationalise decision making on Internet policies by showing true value of media

At the moment, it is difficult to measure the economic value of the creative and media industries since it is hard to track the widespread use on the web. The cultural observatory allows to picture the online use on different platforms and websites.

Project Octopus' interdisciplinary way of understanding the relation between the copyright framework and publishing realities invites to see this relationship in a new light, and to map economic value of cultural flows in a different way.

Ambition 4: Develop a community of provenance information enthusiasts

Existing and abandoned efforts to create a provenance system for the digital commons have been analysed: both where these efforts have been explicitly aimed at creating a provenance system or where the provenance system is a by-product of another service. A main finding in the research paper is that one of the most important aspects why it is difficult to keep a provenance system up-to-date and reliable is because there is no real interaction possible with the data in the system. This project explores new ways of keeping the information in the system in motion by encouraging a strong community to curate the provenance information in a Wikipedia-like manner. The community can interact with the data, check and double-check it.

The repository brings together information from several web silos, aggregators and smaller web environments. This creates a break with the system of having single large silos of information.

Because the observatory adheres to open standards by developing an open technology stack, all data is reusable by others through the API. As such, enthusiasts are invited to come up with even better implementations and use cases of the data, the algorithms and the API.

Ambition 5: Innovate the governance models of provenance systems

The cultural observatory maps cultural flows in new ways and offers a multitude of opportunities to evaluate the (economic) value of cultural works online. As the project draws connections between works that are published online, it also influences other provenance systems and large web silos. It is now possible to understand the influence of provenance systems. This helps to innovate governance models of creative industries online, as well as of provenance systems.

The cultural observatory is open and transparent. As works from several other publishing platforms online are included in the system, these other (closed) systems are urged to be more transparent.

2. Impact

2.1 Expected impacts

Project Octopus removes barriers surrounding web silos (1). This empowers online publishers to explore (the use of) digital media (2). This project supports creative processes. It connects creators and users by establishing a community that documents use of online media (3).

The project also reflects on the role of provenance in the digital market (4). Ubiquitous broadband connections on any device across Europe are just around the corner. It is important to understand the role of provenance information in that digital society. In doing this reflection the project establishes principles and best practices for governance of provenance systems (5).

Ultimately this unleashes the value of latent provenance information on the web (6).

> **Impact 1:** Remove barriers surrounding web silos of media

Walled gardens are closed ecosystems that have internally consistent and well functioning services but restrict access to other web silos. This is a common practice on the Internet. Platforms like Instagram, Facebook and Flickr provide good internal structures for publishing, building upon and attributing other creators. The available provenance information, however, is not readily usable outside of these ecosystems.

Project Octopus removes these barriers by storing provenance information outside of these closed ecosystems. It is no longer required for users to compile provenance information from the location where the media was found. Project Octopus combines provenance sources from multiple media platforms. Information collected by GLAMs and from its community.

An independent community-based and user-oriented provenance repository makes it possible to track a work across the Internet, find reuse permissions and/or find the original creator and publishing platforms. This strengthens the position of the creator and enables the creative reusers in their processes.

> **Impact 2:** Empower Internet publishers to explore (the use of) digital media

The Internet is based on links between web pages and domains. Project Octopus creates a new layer of links that do not normally exist. It connects web pages through the media that is embedded in them. For the first time Internet users gain the ability to explore the Internet

following this flow of distribution. This is a novel way to explore the same work in different contexts.

Furthermore, the system identifies openly licensed works and public domain works. Both types of works can be used by third parties without permission. Works that do not need permission to be used strengthen our digital knowledge society. They, for example, provide important audiovisual context for articles on Wikipedia.

Impact 3: Start a community of Internet users that documents use of online media

Successful communities have large impacts. Wikimedia has transformed the way we deal with encyclopedic knowledge on the Internet. Wikimedia Commons is the largest single collection of reusable media. The upcoming WikiData project revolutionises the way we publish and collect structured data. MusicBrainz is the largest collection of structured data about music. It is their communities that make these projects great, creating a positive spiral of increasingly higher quality data.

Project Octopus relies on strong community partners to kickstart a community of provenance information; and with this community create the same positive spiral.

Impact 4: Improve understanding of the role that provenance systems play in the digital market

Project Octopus' initial research illustrates that there are strong indicators that provenance information is important. The project does legal research into the formal role of provenance system in the digital environment. Combined with a practical workflow in engagement of interested communities as well as making its tools widely available gives a deeper understanding of provenance in our digital society.

The cultural observatory provides a new opportunity to understand this role of provenance systems in the digital society.

Impact 5: Establish principles and best practices for governance of provenance systems

Currently, existing provenance-based systems are either policy-driven copyright registration systems[22] or private identification systems for publishing houses or publishing platforms. Project Octopus establishes principles and best practices from the perspective of Internet

[22] Copyright registration is not in isolation a useful feature. Policy incentives or requirement for obtaining the benefits of some other useful features and a critical mass are needed for registration to have utility.

users and reusers. It takes the dissemination practices into account and focusses on co-creative processes for the creative industries and media industries.

These principles and best practices show a more nuanced perspective than policy-driven copyright registration systems or private identification systems, as their focus is on blocking Internet piracy. It shows the value of provenance information for the creators themselves in their creative processes.

Impact 6: Unleash the value of latent provenance information on the web

Thus far it has been difficult to measure the economic value of cultural works on the web, which is a particularly pressing issue for the creative industries. The cultural observatory visualises cultural flows online, and maps reuse online on several publishing platforms. This enables the calculation of the return of investment on the web. Better investment choices can be made on the outcome of this analysis, in the creative industry as well as on governance level.

Project Octopus allows creators and online publishers to more easily reuse media that they encounter online as it will be possible to derive the provenance of works to figure out the licensing status. It becomes easier to ask for permission to reuse works if works are not freely available. This can be expected to increase the number of economic transfers on the web.

Creators and rights holders get access to a tool that shows where their works are used online. This helps them understand the dissemination of their works. In case of copyright infringement the creators and rights holders can also take more effective actions.

2.2 Measures to maximise impact

Project Octopus takes a broad perspective on digital provenance systems, it combines state-of-the-art technology development with excellent research and has highly successful community partners to develop engagement and maximise impact of the project.

Project Octopus delivers a community-based cultural web observatory that enables its users to trace publishing contexts of a creative work. The projects produces three kinds of results: Technological innovations, research and policy papers, and a community of provenance enthusiasts.

2.2.1 Measures to maximise exploitation

Exploitation is ensured as all the project's products are published as openly as possible (1). All technological innovations are published on collaboration tools that are open (2), and so are the

results of Octopus' research (3). The observatory is designed to engage with a strong audience to keep information up to date (4).

Measure 1: Publish as openly as possible

The project publishes all of its products as open content, open source, open data and open formats to minimise barriers for reuse outside of the project and after the end date of the project. All its source code is published using a European Union Public Licence v.1.1. license. Its written products are licensed using a Creative Commons Attribution 4.0 license. Finally, its collected data is made available under a Creative Commons Zero Public Domain Dedication 1.0.

This publication policy allows third parties to engage with the project results. Others can adopt the standards of the project's results. Third parties can also reuse it for commercial purposes, building upon this work by a FLOSS community and crossover effects in other sectors.

The project develops its tools API-first, this means that interaction between Project Octopus and other services via machine to machine interfaces is the first architectural principal. Project Octopus is not only open in the sense of intellectual property rights (e.g. open licenses), but its services also invite others to participate.

Measure 2: Use open collaboration tools

Participation and adoption outside of the project is possible. By adopting GitHub[23] as our source code repository, interested third parties are able to comment, contribute and build upon the work delivered in the project. The common use of GitHub makes this a credible path to bring the technological innovations to the market.

GitHub can store the project's source code after the project has ended. As a free third party service it ensures that products of the project will not be locked up in a closed repository at of the partners of the project. This ensures the durability of the project.

Furthermore, the project aims to develop media recognition tools that can be run in a web-browser and on servers with a single codebase. This creates the opportunity to distribute necessary processing powers to users and enables all internet users to adopt the developed tools for their own projects and goals.

[23] A commonly used web-based source code repository hosting service that allows for open collaboration between partners and third parties.

Measure 3: Publish research products as gold open access

All research and research data are published openly and free for reuse. The project applies a Creative Commons Attribution 4.0 license to all reports, white papers, and publications. To ensure maximum reuse and dissemination of research products, legal research are published in gold open access journals.

Measure 4: Strong community engagement

Public data gathering relies on contributions by many different parties. It is an intrinsic part of the project to build a strong community of individuals and institutions that actively engage with provenance information in the repository. Several consortium partners bring a strong network to the project (WMDE, ToN, IR, BnL). All these partners have their expertise in different media types (audio, audiovisual, images). It is their task to reach out to an even wider community to create a group of users that actively engage with provenance information in the repository.

2.2.2 Measures to maximise communication

Project Octopus works in close cooperation with creative communities, GLAMs and remix communities throughout. It internalises research and needs of its user groups (5). Project Octopus addresses those groups specifically by organising events (6). A symposium cuts through the different disciplines and brings together interested parties from different sectors and backgrounds (7). Finally a project website and an active social media policy inform the world about the project and its goals (8).

Measure 5: Thorough research in all decision-making aspects

Project Octopus has many interdependencies and interrelations. Scholarly research and community research are set up as feedback systems for the specific development tasks to ensure that research forms the backbone of all decision-making.

This includes research into the relation of provenance systems and the copyright framework, research into privacy issues when collecting latent provenance information from around the web and gathering specific requirements and needs from three highly specialised user groups (creative communities, GLAMs, remix communities) of three different media types (audio, audiovisual, images).

Measure 6: Events

Project Octopus organises seven community events for specific communities (creative communities, GLAMs, and remix communities). The events are not stand-alone events but reach out to specialised communities during events that these sectors themselves organise. This ensures maximum impact and reach.

Measure 7: Symposium

Project Octopus' symposium brings together experts from a variety of fields, from technologists to (legal) researchers, GLAMs to ambassadors of remixers and creative users. All these experts deal with issues surrounding provenance systems on a daily basis. The symposium marks the finalisation of the initial development stage as the core infrastructure will be released as a beta version during this event. It provides the first moment of feedback on this core infrastructure.

The symposium is simultaneously the kick-off of the community phase of the project. Users are invited to start using the cultural observatory and to start curating information.

Measure 8: Project website and outreach on social media

The project website is designed to keep both consortium members and other interested parties updated about the state of affairs of the project. It collects all products, provides regular updates for interested parties and serves as a landing page for background information. Social media is used to reach out to a wider audience. Project Octopus has an active social media policy. A visual identity – such as a recognisable logo – helps to make Project Octopus indistinctly recognisable to audiences and communities.

3. Implementation

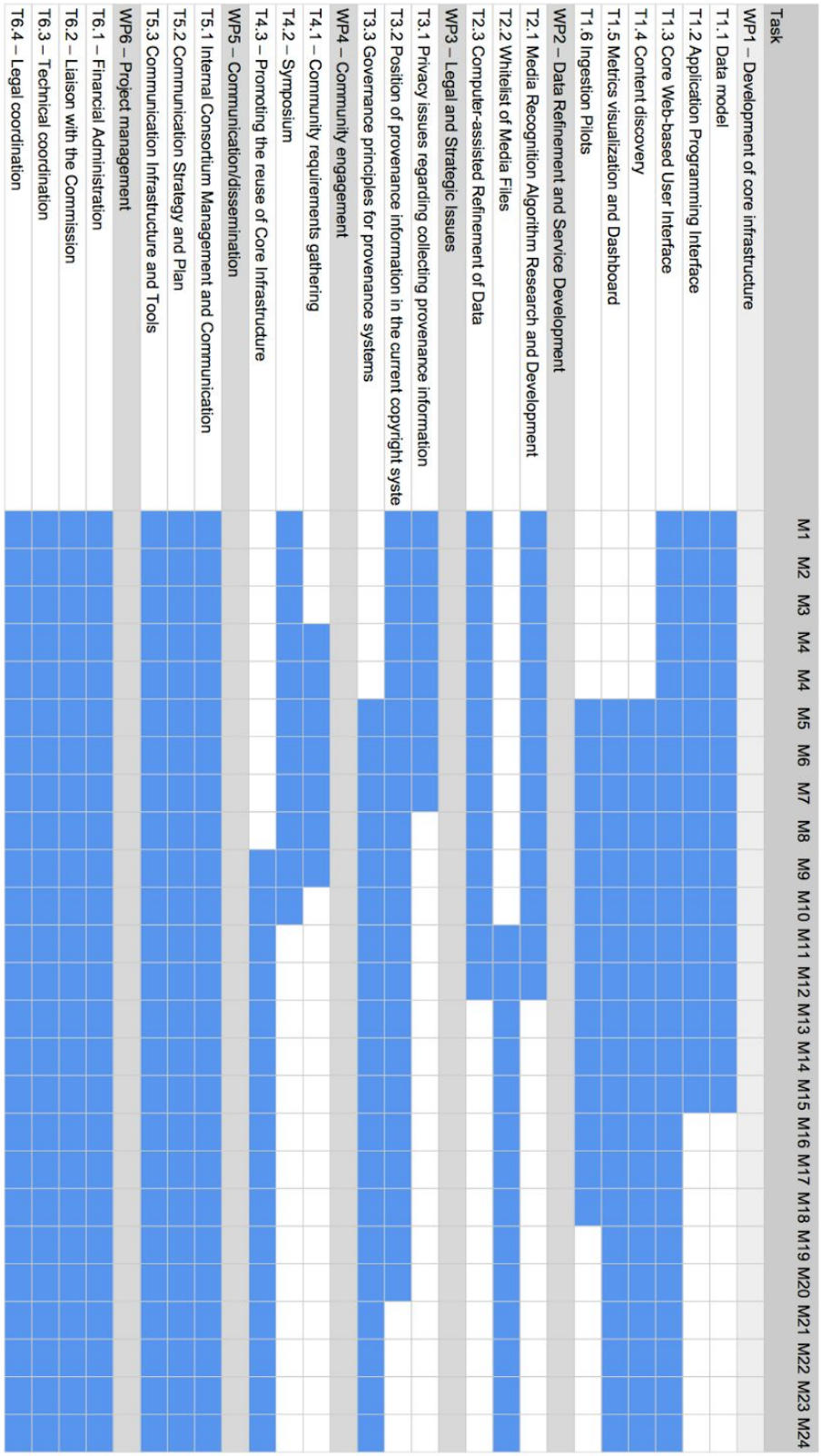
3.1 Work plan — Work packages, deliverables and milestones

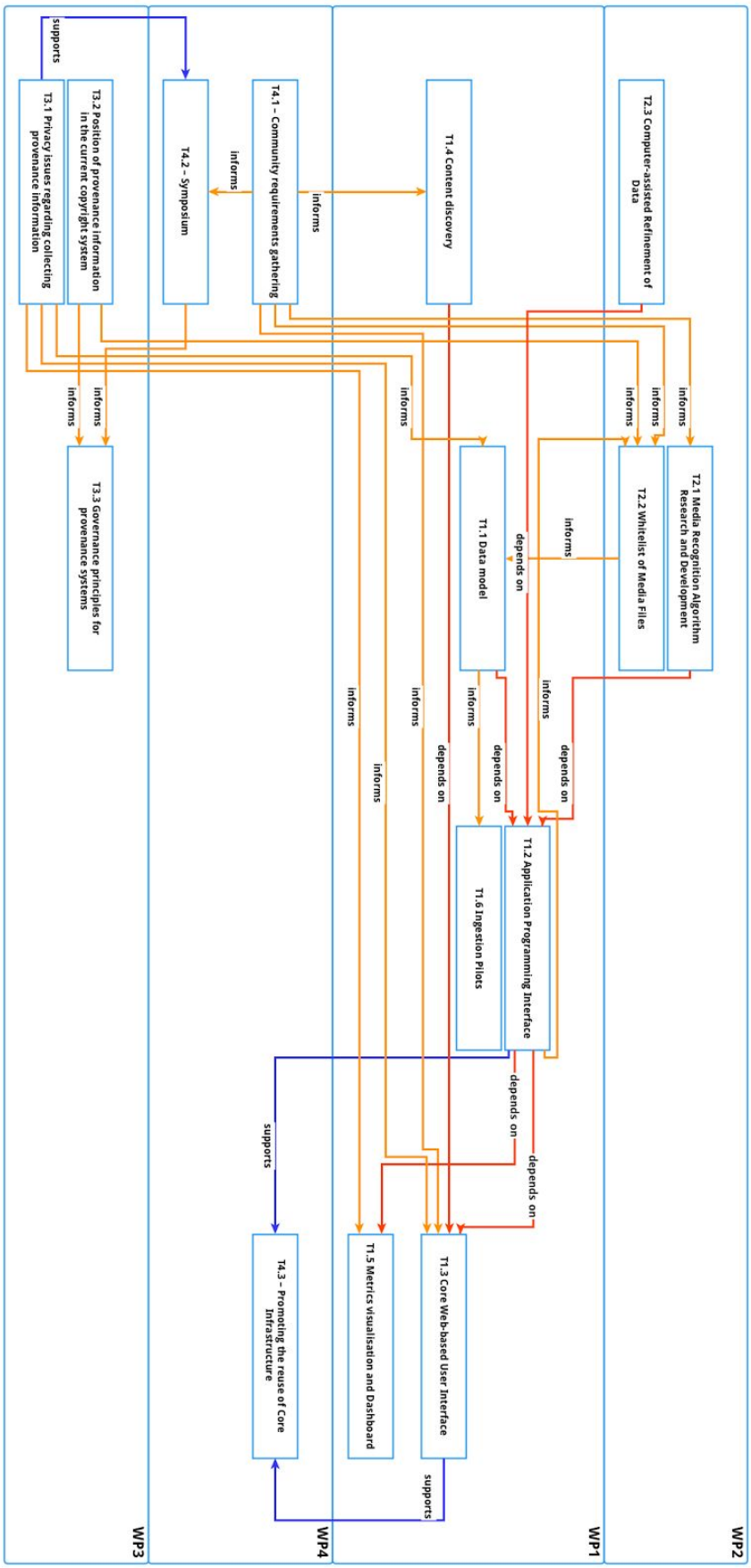
The work plan consists of 6 work packages. Each with a domain expertise necessary for the project. They contain technological development (WP1, WP2), scholarly research (WP3), community engagement (WP4), communication and dissemination (WP5) and project management (WP6):

- ➔ **Work Package 1 – Development of Core Infrastructure** develops the main technological parts of the project. It researches data models, requirements of user groups, and develops interactions between services and this core infrastructure.
- ➔ **Work Package 2 – Data Refinement and Service Development** enables new services to be build on top of the core infrastructure developed in WP1. These services include data refinements and whitelists creation.
- ➔ **Work Package 3 – Legal and Strategic Issues** researches the legal and strategic issues. It aims to provide strategies for dealing with legal issues arising from the collection, organisation and dissemination of provenance information (such as privacy, data quality and accuracy), as well as the sustainable governance of the system and its infrastructure.
- ➔ **Work Package 4 – Community Engagement** reaches out to three types of communities (creative communities, GLAMs, and remix communities). It gathers the needs of these user groups. After a stable initial version of Octopus has been released the participants of this work package reach out to these user groups and introduce the developed tools.
- ➔ **Work Package 5 – Communication/Dissemination** facilitates efficient communication between partners, and creates the infrastructure to guide external communication and project outreach.
- ➔ **Work Package 6 – Project Management** is responsible for the management of the whole project, according to agreed methods, structures and procedures.

WP No.	Work Package Title	Lead No.	Lead Short Name	Person-Months	Start month	End month
1	Development of Core Infrastructure	2	PP	50	1	24
2	Data Refinement and Service Development	3	CM	34	1	24
3	Legal and Strategic Issues	5	IViR	39	1	20
4	Community engagement	4	IR	46	1	24
5	Communication/ Dissemination	1	KL	14	1	24
6	Project Management	1	KL	24	1	24
				207		

The work plan involves two years with nine partners. Its project planning is presented in the GANTT chart on the right.





The interrelations between the tasks described above in the GANTT chart are visually displayed in the figure on the left. They are categorised in tasks that support other tasks (purple), tasks that inform other tasks (yellow) and tasks that depend on other tasks (red). Work packages 5 and 6 are not displayed in this figure. They support all tasks in the project.

3.1.1 Work package 1

Work package number	1		Start Date or Starting Event				1
Work package title	Development of Core Infrastructure						
Participant number	2	3	1	8	4	6	7	9
Short name of participant	PP	CM	KL	BnL	IR	KT	ToN	WM DE
Person/months per participant:	27	10	3	1	1	4	1	2

Objectives

The objective of this work package is to develop the core infrastructure of a cultural observatory that ingests provenance and usage information from the web, repositories, and creators, and is curated by the community. It results in the availability of reliable documentation of the provenance of works, relationships among works, uses of works, and new ways to discover culture and understand cultural flows. The core infrastructure provides an API which can be used by third parties (and extensively by WP2) to ingest provenance and usage information from repositories and to add value to the system through extended functionalities.

Description of work

This work package consists of the following tasks:

- ➔ T1.1 Data model: develops the data model for storing, updating, and querying provenance and usage information.
- ➔ T1.2 Application Programming Interface (API): develops an API that provides a machine interface for adding, updating, and querying provenance and usage information.

- ➔ T1.3 Core web-based user interface: develops a website to view, add and update provenance and usage information.
- ➔ T1.4 Content discovery: develops discovery functionality for the web user interface.
- ➔ T1.5 Metrics visualisation and dashboard: develops a dashboard for visualisation of metrics about collections of works.
- ➔ T1.6 Ingestion pilots: ingests data from large data providers into the database.

Each task is elaborated in the section below.

T1.1 Data model (M1-15)

Lead: PP, Participants: CM, KL

This task develops the data model for storing, updating, and querying provenance and usage information. As there are multiple ways of storing and maintaining provenance information for different media types, this task researches current implementations and collaborates with other project partners to integrate existing frameworks.

The data model is implemented in a database which forms the core of the Octopus web observatory, supporting:

- ➔ Ingestion of provenance and usage information for works and digital objects from repositories and the web.
- ➔ Assertions about privileged relationships that agents have with a work (a superset of copyright registration).
- ➔ Descriptions of relationships among works.
- ➔ Authenticated users.
- ➔ An auditable and immutable recording of updates of provenance information.
- ➔ The ability to search and create aggregate metrics for works in the system.

Additionally the data model supports storing and querying media recognition information developed in T2.2 and information pertinent to rights status determination in T2.2.

A description of the design and rationale of the data model and the schema for a database implementing this model is part of D1.1.

T1.2 Application Programming Interface (API) (M1-15)

Lead: PP, Participants: CM, KL

This task develops an API that provides a machine interface for adding, updating, and querying provenance and usage information held in the database built in T1.1. The API is the primary way for programs implementing both core user interfaces and extended functionalities to interact with the database.

A description of the API is part of D1.2.

T1.3 Core web-based user interface (M1-24)

Lead: PP

This task develops a website to view, add and update provenance and usage information. Provenance information contributed from many sources can be difficult to understand and to validate. Therefore, this task includes research, testing, and refinement of user interfaces for interacting with provenance information. The user interface encourages the public as well as entities with authoritative relationships with works to contribute and improve provenance information, and make the history of updates of provenance information accessible.

The user interface utilises the media recognition capabilities developed in T2.1 to assist provenance curation interactions.

A description of the user interface and underlying research is part of D1.3.

T1.4 Content discovery (M6-24)

Lead: PP

This task develops discovery functionality for the web user interface. This functionality is grouped into three categories: 1) query-based discovery (find works based on text and attribute search), 2) browsing by system and user-defined lists (e.g., works with specified attributes, such as a particular genre and rights status, ordered by recency and popularity), and 3) enhancements to the provenance viewing and curation user interfaces (e.g., showing related works).

A description of discovery functionalities is part of D1.4.

T1.5 Metrics visualisation and dashboard (M6-24)

Lead: PP

This task develops a dashboard for visualisation of metrics about collections of works, both internal system metrics (e.g., curation activity around works in collection) and external use metrics based on relationship and usage information documented by the system (e.g., timeline of reuses). While dashboards for arbitrary collections of works are possible, the two main groupings to be supported are works claimed by a creator, and works in the collection of an institution, in each case giving the interested party an unprecedented view into the use of their works across the web.

A description of dashboard functionalities is part of D1.5.

T1.6 Ingestion pilots (M6-18)

Lead: CM, Participants: PP, KL, BnL, IR, ToN, WMDE

This task develops three pilots to explore and execute different ways of ingesting provenance information into the database. It includes researching data models of other stores of content like Wikimedia Commons, Europeana, Flickr, etc. and converting these to the data model described in T1.1, and pilot ingestion implementations corresponding to the three user groups in WP4. The implementations are reusable components for future ingestions and different partners.

A description of the three pilots for ingestion is part of D1.6.

Milestones

- ➔ M1.1 Initial user interaction research complete (M3).
- ➔ M1.2 Initial data model, initial database schema reflecting model; database up and running; updated API design and initial implementation; initial design and user interaction implementation (M6).
- ➔ M1.3 Database schema and API updated based on feedback from T1.3 implementation; user interface implementation uses API backend; API ready for beta users; initial content discovery and dashboard designs; research on external data models for ingestion complete (M9).
- ➔ M1.4 Database schema, API and user interface updated based on feedback from T1.4 and T1.5 implementations; initial content discovery and dashboard implementations; initial ingestion of provenance information from all pilots (M12).
- ➔ M1.5 Database schema and API updated based on feedback from T1.6 and WP2; ingestion from pilots complete (M15).
- ➔ M1.6 User interface, content discovery and dashboards updated based on feedback from T2.3 and T4.3 (M24).

Deliverables

- ➔ D1.1 Online documentation of the Octopus data model research and design (M9).
- ➔ D1.2 Online documentation and initial implementation of the Octopus API research and design (M10).
- ➔ D1.3 Uniform user interfaces available based on a corporate identity, user interface research and design (M11).
- ➔ D1.4 Discovery features implemented on Project Octopus (M14).
- ➔ D1.5 Metrics features implemented on Project Octopus (M16).
- ➔ D1.6 Ingestion pilots finished from three different sources (M18).

➔ D1.7 Core infrastructure complete, stable, source available (M24).

3.1.2 Work package 2

Work package number	2		Start Date or Starting Event			1	
Work package title	Data Refinement and Service Development						
Participant number	3	6	1	8	2		
Short name of participant	CM	KT	KL	BnL	PP		
Person/months per participant:	17.5	10.5	3	1	2		

Objectives

The objective of this work package is to explore how the information in the database can be curated to add value to it, beyond the information initially imported into it. The work package explores three types of added value to the database: media recognition, whitelisting, and computer-assisted refinement of data. In each case, autonomous agents interact with the database through its API, retrieving information relevant for its processing, and then feeding the results back into the database as additions or changes, thereby contributing to gradually refine the information in the database and opening up the possibility of using this information for better services and tools.

Description of work

This work package consists of the following tasks:

➔ T2.1 Media recognition algorithm research and development: researches and develops open algorithms to determine similarities between media files.

- ➔ T2.2 Whitelist: researches and implements methodologies to determine the rights status of creative works.
- ➔ T2.3 Computer-assisted refinement of data: develops the processes and software needed for communities of users to create autonomous agents that interact with the database in order to refine the information in it.

T2.1 Media recognition algorithm research and development (M1-12)

Lead: KT, Participants: CM

This task researches and gathers requirements for algorithms that can be used to evaluate similarities between two media files: video, audio and images. This helps to recognise media that has been slightly transformed, derived from another work, or stripped of provenance data, so that this information can then be included in the database. This makes desired functions of a provenance store more effective or efficient.

A series of existing algorithms is studied and assessed on their level of usefulness for the purpose of our project. The algorithms are assessed using a representative sample of creative works from online sources under an open license. Whether these algorithms are protected by patents is checked in T6.4. The project only uses algorithms that are not protected by patents and aims to publish new implementations as open standards and open source to ensure maximum reusability of the project's outcomes.

These recognition algorithms are developed in programming languages that are suitable for both server and client applications. Algorithms implemented for the sole purpose of running on the servers that host the provenance store can lead to a computing bottleneck. Enabling end users to create their own list of identifiable aspects of media files leads to a distributed method of media recognition that will not be as costly for the project.

An overview of the algorithms for media file recognition is part of D2.1, and the complete documentation of the algorithms implemented is part of D2.2.

T2.2 Whitelists (M1-24)

Lead: KL, Participants: CM, BnL

This task researches possibilities for developing whitelists with the intention to create clarity about the rights status of media files. The task researches the requirement for a service that enables the determination of the public domain status of a (copyright or related rights protected) work and communicates necessary changes to the project data model developed as part of M1.5.

Outcomes of the rights determination services are fed back into the WP1 developed database through its API. A whitelist service is available through the user interfaces developed as part of M1.6.

There is close collaboration with Project OutOfCopyright.eu, an open service for rights determination developed by project partners KL, IViR, and BnL. This task combines the APIs of that platform with the APIs of Project Octopus.

The requirements for whitelisting, together with its implementation, are documented and delivered in D2.3.

T2.3 Computer-assisted refinement of data (M12-M24)

Lead: CM, Participants: PP

This task establishes and develops the processes and software needed for communities of users to create autonomous agents that interact with the database in order to refine the information in it. Such agents could use machine learning, algorithms or leverage the power of a crowd of users in order to determine the changes needed to increase the quality of the metadata in the database. This happen in an iterative process where multiple agents interact with the database to form a continuous improvement and learning cycle. Informed by T4.1, this task leads to the creation of three to five autonomous agents that each represent a particular need of the target community identified in WP4.

Milestones

- ➔ M2.1 Create a representative sample of media files (M1).
- ➔ M2.2 Perform initial assessment and validation of algorithms (M3).
- ➔ M2.3 Implement, document and demonstrate algorithms for end-user use (M6).
- ➔ M2.4 Ascertain minimum viable information required of the data model for whitelisting (M3).
- ➔ M2.5 Implement a whitelist in Outofcopyright.eu using data from the project (M12).
- ➔ M2.6 Identify tasks suitable for autonomous agents (informed by T4.1) (M12).
- ➔ M2.7 Implement three autonomous agents and demonstrate against the project database (M18).

Deliverables

- ➔ D2.1 Overview of suitable openly available algorithms to support the discoverability features of the project (M6).

- ➔ D2.2 Implementation and documentation of the algorithms implemented in the project (M12).
- ➔ D2.3 Report on data requirements and implementation of whitelisting of media files for public open reuse (M4).
- ➔ D2.4 Processes defined and implemented for continuous refinement of data with autonomous agents (M22).

3.1.3 Work package 3

Work package number	3	Start Date or Starting Event			1
Work package title	Legal and Strategic Issues				
Participant number	5	1	4	2	8	3
Short name of participant	IViR	KL	IR	PP	BnL	CM
Person/months per participant:	17	7	7	4	1	3

Objectives

This work package deals with legal and strategic issues that we expect to arise as a consequence of building a repository of provenance information for cultural objects on the web. The work package aims to provide strategies for dealing with legal issues arising from the collection, organisation and dissemination of provenance information (such as privacy, data quality and accuracy), as well as the sustainable governance of the system and its infrastructure. An open repository of provenance information for cultural objects, as it is proposed in this project, can be expected to resolve a number of issues that hamper the functioning of copyright in the digital environment. This work package explores how a registry of provenance information can improve existing copyright licensing and rights clearance practices and propose future use cases based on this research.

Description of work

This work package consists of the following tasks:

- ➔ T3.1 – Privacy issues regarding collecting provenance information: Research into the privacy implication of harvesting, aggregating and storing provenance information in an open online repository for provenance information.
- ➔ T3.2 – Provenance information in the context of the existing copyright framework: Research into the relationship between an open online provenance information repository and the existing legal framework in the field of copyright
- ➔ T3.3 – Governance principles for provenance systems: Research into best practices for governing an open online provenance information repository including research mechanisms to ensure the accuracy and reliability of such systems.

T3.1 – Privacy issues regarding collecting provenance information (M1-7)

Lead IViR, Participants: KL, CM, IR

Harvesting, aggregating and storing provenance information in an open online repository can be expected to have a number of privacy implications. Provenance information is by definition information that is related to individuals and specific legal entities. While collecting information about the links between works and their creators/owners has a number of highly desirable consequences that form the basis of this proposal, it also means that such a system inevitably deals with the collection and processing of personally identifiable information. In this task the IViR with input from PP, KL and IR map the personal data protection issues that can be expected to arise from the project and develop strategies and guidelines for dealing with these issues from a European perspective.

The work package produces a set of design principles that feeds into work undertaken as part of WP1 (specifically T1.1, T1.2, T1.3 and T1.5) and inform the personal data protection policy that is being developed in T6.4

T3.2 – Position of provenance information in the current copyright system (M1-20)

Lead IViR, Participants: PP, IR, KL, BnL

While the usefulness of provenance information is not only limited to copyright related questions, it is evident that a well-functioning open online provenance repository has the potential to improve the way copyright is exercised in the online environment and the way copyrights are cleared. In order to realise this potential of Project Octopus this task needs to identify benchmark criteria according to which the collected provenance information can be deemed reliable enough to be used in licensing and rights clearance contexts. In addition this work package researches the specific requirements of use cases that place provenance in the context of copyright (whitelist, improving open content

licenses, conflict resolution) and develop criteria based on this research. This work package also undertakes research that explores the relationship between provenance information and copyright formalities, other types of rights management information and existing private enforcement systems (such as contentiD).

T3.2.1 Provenance information as a driver for more efficient licensing and rights clearance

This subtask looks into the question how an open online provenance repository can improve rights clearance operations (via a whitelist of freely reusable works or otherwise) and how it can improve existing open licensing systems (such as Creative Commons) that currently suffer from weak links between the licenses, licensors and the licensed works. Research is carried out to better understand these problems and suggest benchmark criteria for use cases that build on top of Project Octopus. The results of this subtask feed into T2.1 and T2.1

T3.2.2 Legal validity of provenance information in the context of the copyright framework

This subtask researches the requirements on a provenance system that results from the need to use this information in the context of the copyright framework. How does the registration of provenance information relate to the prohibition of formalities established by the Berne Convention? How can conflicting claims of ownership be reconciled with universal truth assumption that underpins the concept of authorship inherent to copyright? What are the legal consequences of claiming ownership or making other assertions in relation to specific works? How can an open provenance repository interact with proprietary rights management systems such as ContentID and the information held in the databases of collective rights management organisations? These questions are examined in the context of the ongoing development of Project Octopus with the objective of supporting design decisions. The outcomes of this research is also crucial in establishing support for our approach among other stakeholders within the creative industries.

T3.3 – Governance principles for provenance systems (M6-24)

Lead KL, Participants IViR, PP, CM, IR

This task develops governing principles for provenance systems that feed into the the governing model for Project Octopus. Project Octopus is not developed as a proprietary database and as a result it requires a governance system that both ensures the quality of the data information that is provided by the system and the openness of the system. As part of this task we first develop and publish principles for open web-based provenance systems. These principles are discussed with a wide variety of stakeholders during the symposium that is being organised as part of T4.2. Based on these principles we then

develop a governance model for Project Octopus that can contribute to its sustainability after the end of the project period. As part of this task we also develop criteria for ensuring an adequate level of reliability of provenance information that can serve as a basis for the exchange of provenance information between Project Octopus and other provenance information repositories.

Milestones

- ➔ M3.1 Map personal data protection issues expected to arise during the implementation of Project Octopus (M2).
- ➔ M3.2 Design principles to ensure compliance of Project Octopus with personal data protection rules (M5).
- ➔ M3.3 Map potential interactions between open provenance systems and copyright licensing and rights clearance practices (M8).
- ➔ M3.4 Formulate initial principles for open web based provenance systems (M10).
- ➔ M3.5 Define provenance reliability principles and models (M10).

Deliverables

- ➔ D3.1 Research report on personal data protection issues related to open online repositories for provenance information (M7).
- ➔ D3.2 Research report on Provenance information as a driver for more efficient licensing and rights clearance (M12).
- ➔ D3.3 Research report on the Legal validity of provenance information in the context of the copyright framework (M20).
- ➔ D3.4 Governance model for Project Octopus (M24).

3.1.4 Work package 4

Work package number	4		Start Date or Starting Event			1
Work package title	Community Engagement					
Participant number	4	8	7	9	1	5
Short name of participant	IR	BnL	ToN	WMDE	KL	IViR
Person/months per participant:	16	6	12	9	2	1

Objectives

The aim of this work package is twofold: First it gathers the needs of communities for provenance information to be fed back to WP1 and WP2, and after M10 it creates outreach about the products .

For this purpose, the work package engages communities and individuals that either work with provenance information, use provenance information and/or rely on provenance information. Three user groups are identified: GLAMs, creative communities and remix communities.

The second part of this project starts after a stable initial version of Octopus has been released in M10. It reaches out to the three user groups and introduces the developed tools. Initially by organising an invite-only symposium on provenance information and then ensuring wide dissemination and promoting engagement of target communities. The project makes use of the networking, clustering capacity and multiplicator effects of partner IR, innovative GLAM institute the BnL, the creative community that backs ToN, and the remix communities active in Wikipedia and other Wikimedia projects.

Description of work

This work package consists of the following tasks:

- ➔ T4.1 – Gathering of community use cases and societal needs to inform Octopus development. Gathering and reporting on use cases to inform Octopus development.
- ➔ T4.2 – Symposium - community building and feedback. Organising a symposium for network building and feedback.
- ➔ T4.3 – Reuse and Feedback Core Infrastructure. Sustaining the adoption of the Octopus core infrastructure through continuous dialogue with user network.

Each task is elaborated in the section below.

T4.1 – Gathering of community use cases and societal needs to inform Octopus development (M1-9)

Lead: IR, Participants: BnL, ToN, WMDE

This task identifies and brings together a group of potential partners and beneficiaries of the Octopus system. It focuses on outreach and informs tasks T1.3-6, T2.1, and T2.2. The task identifies problems that user groups have with regards to provenance information in their respective uses of the digital environment.

Task 4.1 consists of three subtasks, each focussed on identifying the needs and problems of a specific user group regarding their use, or lack of use, of provenance information in their respective digital environments. Subtasks 4.1.1 (GLAMs), 4.1.2 (Creative Communities) and 4.1.3 (Remix Communities), led by BnL, ToN and IR resp. Each partner draws upon a large existing network in its user group to make sure that the gathered information offers the best possible input for the Octopus system. Each user group deals with a specific set of problems relating to provenance information. The task informs the project about these specific needs.

This task:

- ➔ Identifies and builds a network of partners, using existing networks, channels and events. Once the network is gathered, its members collaboratively work together using online tools.
- ➔ Identifies needs and problems of different user groups with regard to provenance information: staff of GLAM institutions, creators from a wide variety of fields (music, photography, journalism, filmmaking and more), users of Wikimedia Commons, Flickr and the like, app developers, editors of Wikidata and others.

➔ Produces one report for each user group (GLAMs, Creative Communities, Remix Communitie) identifying their needs to be fed back into the development process.

The results of this task are described in D4.1 Report of feedback from user groups. (M10).

T4.2 – Symposium - community building and feedback. Organising a symposium for network building and feedback (M1-10)

Lead: IR, Participants: KL, PP, CM, IViR, KT, ToN, BnL, WMDE

This task organises a two-day symposium that brings together technologists, (legal) researchers, users (community) and advocates to discuss technological solutions to provenance and registration of works. The symposium takes place in M10 coinciding with the public stable release of the core infrastructure (M1.3) and the kickoff of the community-use phase of the project. Participants are introduced to the core infrastructure and invited to comment and discuss its features. The symposium serves both as a community building event, bringing together the network established in T4.1, and as a feedback mechanism, laying the groundwork for further collaboration in T4.3.

The results of this symposium are made available through D4.2 Online proceedings of symposium on provenance information. (M12)

T4.3 – Promoting the reuse of core infrastructure sustaining the adoption of the Octopus core infrastructure through continuous dialogue with user network (M10-24)

Lead: IR, Participants: BNL, ToN, WMDE

This task brings the Octopus tool to our three user groups (GLAMs, Creative Communities, Remix Communitie). We identify relevant stakeholders and provide them with information on how to use and benefit from the Octopus system. The user groups are, on the one hand, invited to explore the tool itself, discover its possibilities and functionalities, and on the other hand, contribute to the information on Octopus by curating, editing and adding provenance information. For a large outreach, user guidelines for information crowdsourcing are developed and made available online.

Users of the tool can give feedback on the tool itself to improve it and create valuable user experiences. This feedback improves the core infrastructure and user interface built in WP1 and WP2. As such, a feedback mechanism for testing conceptual as well as technological iterations of the development (set-up, composition, usability etc.) is developed, and this feeds results back to the development team.

The task aims to be responsive to feedback from an audience as broad as possible. The feedback process has several iterative loops. Initial needs are gathered from our user groups during the events organised in T4.1.

The products of this task feed into D4.3 User guidelines for use and participation in Project Octopus (M14).

Subtasks 4.3.1 (GLAMs), 4.3.2 (Creative Communities) and 4.3.3 (Remix Communities), led by BnL, ToN and IR respectively promote the system to user groups

The online collaboration tools established in T4.1 and promoted at the symposium remains the backbone of this continued outreach. For all user groups, "push" formats, i.e. in form of electronic newsletters, continuously serve to keep stakeholders up to date on important improvements and advances of Octopus. A permanent feedback mechanism is offered via the web.

Additionally, other venues for outreach are used to reach new and additional potential users. These venues are adapted to each user group:

For Creative Communities, ToN engages with a broad spectrum of stakeholders from all relevant fields of creative industries. ToN achieves this by organising meetups during relevant events for musicians, game developers, bloggers, film-makers and others (i.e. re:publica, MIDEM, mipcom, Game Developers Conference). There, potential ambassadors for future outreach in each sector are to be identified who then continue to interact with the respective communities.

The GLAM sector is dealt with by BnL using a two-pronged approach: The relevant IT /innovation communities are relatively compact and well connected, plus there are the very large Europeana Network and EuropeanaTech communities. Most of these users community will in all likelihood be identified and networked in T4.1 and T4.2. In T4.3, BnL identifies key active members of the Octopus network who then, as ambassadors, themselves promote Octopus at diverse events (such as Code4Lib, Computers in Museums, etc.) and thus drive more participants to the online collaboration tools. A different approach is needed for the large software vendors who dominate the library, archives and museum market, including specialist providers such as, for example, the the IIPC (International Internet Preservation Consortium) because the development cycles are slow and there are no events that bring together the GLAM sector software vendors and could be piggy-backed upon. Therefore such players are approached individually and invited to a workshop aimed at exploring integration options for Octopus.

IR organises, in cooperation with WDME, workshops and meetups for the Remix Community, including Wikimedia projects (i.e. Wikimedia Commons) and chapters from various EU countries, Creative Commons country organisations, and more.

The products of these activities will be documented in D4.4 Report on outreach (M24).

Milestones

- ➔ M4.1 Report on Creative Communities (M6).
- ➔ M4.2 Report on Remix Communities (M6).
- ➔ M4.3 Report on GLAMs (M6).
- ➔ M4.4 Continuous documentation of feedback to development in online collaborative too (M6, M10, M14, M18).
- ➔ M4.5 Symposium (M10).

Deliverables

- ➔ D4.1 Report of feedback from user groups (M10).
- ➔ D4.2 Online proceedings of symposium (M12).
- ➔ D4.3 User guidelines for use and participation in Project Octopus (M14).
- ➔ D4.4 Report on outreach (M24).

3.1.5 Work package 5

Work package number	5			Start Date or Starting Event			1
Work package title	Communication/Dissemination						
Participant number	1	2	4				
Short name of participant	KL	PP	IR				
Person/months per participant:	10	2	2				

Objectives

This work package facilitates efficient communication between partners, and creates the infrastructure to guide external communication and project outreach. Its focus is on

developing products that are helpful in the process of outreach to the wider public. Part of the external communication plan are the development of a project website and templates fit for the project. All these products are consistent with the design of Octopus, thus help to create a recognisable end product that the project delivers to the wider public.

Description of work

This work package consists of the following tasks:

- ➔ T5.1 Communication.
- ➔ T5.2 External communication strategy and plan.
- ➔ T5.3 Communication infrastructure and tools.

Each task is elaborated in the section below.

T5.1 Internal communication (M1-24)

Lead: KL, Participants: PP, IR

This task takes care of internal communication with all consortium partners. It is ensured that all decision-making complies with beforehand agreed to agreements. The task leader organises two meetings with all consortium partners: a project kick-off meeting and a meeting in the second year of the project. It further takes care of the organisation of executive board meetings and provides secretarial support.

T5.2 External communication strategy and plan (M1-24)

Lead: KL

This task develops and implements a strategic plan for communication and dissemination for Project Octopus. Focus is on key communication objectives to reach the relevant target audiences and stakeholders identified in WP4 (GLAMs, remix communities and creative communities), as well as potential users to add and edit provenance information.

The plan is updated continuously throughout the project. It keeps track of the different communication activities that are set out in the description of work.

T5.3 Communication infrastructure and tools (M1-24)

Lead: KL, Participants: PP

This task is primarily responsible for setting up the infrastructure for external communication. This includes facilitating platforms and producing publicity materials necessary for dissemination. For this purpose, a visual identity is developed such as a project logo. This is used on all products that are disseminated. There is outreach to a wider audience by use of social media. The strong potential of social media for exposure of the project is realised by setting up relevant social media channels to distribute information, updates and content.

The task develops a project website that is designed to take care of project organisation. The project website contains detailed information on the aims, objectives, consortium, work processes and the current state of Project Octopus. It provides information for all interested parties and the general public. This website hence differs from the Project Octopus user interface.

Decisions as part of T5.3 are taken in close cooperation with the coordinators of WP4 (IR, ToN, WMDE, BnL).

Milestones

- ➔ M5.1 Kick-off meeting (M1).
- ➔ M5.2 All partners meeting (M14).
- ➔ M5.3 Visual identity for Project Octopus (M4).

Deliverables

- ➔ D5.1 Strategic plan for communication and dissemination (M3).
- ➔ D5.2 The project website (M6).

3.1.6 Work package 6

Work package number	6			Start Date or Starting Event		24		
Work package title	Project Management							
Participant number	1	2	3	8	7	9	4	5	6
Short name of participant	KL	PP	CM	BnL	ToN	WMDE	IR	IViR	KT
Person/months per participant:	9	6	3	1	1	1	1	1	1

Objectives

This work package is responsible for the management of the whole project, according to agreed methods, structures and procedures. This includes administrative management, reporting to the EC, performance monitoring and technical coordination. A set of specific progress reports are produced as part of good practice in project management.

Description of work

This work package consists of the following tasks:

- ➔ T6.1 – Financial administration.
- ➔ T6.2 – Liaison with the Commission.
- ➔ T6.3 – Technical coordination.
- ➔ T6.4 – Legal coordination.

T6.1 – Financial administration (M1-24)

Lead: KL, Participants: PP, CM, IR, IViR, KT, ToN, BnL, WMDE

The coordinator is responsible for the financial administration. The coordinator receives project finance from the EC, distributes it to the partners in time and maintains necessary records. The coordinator also monitors the budget and provides report at quarterly

intervals. If necessary, specific approval of budget changes are sought from the EC. The coordinator is responsible for submission of the financial statements throughout the project.

T6.2 – Liaison with the Commission (M1-24)

Lead: KL

The coordinator is in contact with the Commission, and is in charge of any formal documentation that ought to be provided by participants. At the same time, the coordinator makes sure to oversee that generic issues such as gender equality are taken care of. The coordinator accumulates annual Progress reports, with input from all work packages.

T6.3 – Technical coordination (M1-24)

Lead: PP, Participants: CM, KT, KL

The technical lead provides ongoing technical overview and direction for the project, enforcing a common approach and bringing the technical teams together during each phase of the development, to reach a function-complete, tested state. The technical lead facilitates technical communication between project participants, participating in the planning, development and testing process.

The Data Management Plan (DMP), as indicated in the proposal form, is written by KL.

T6.4 – Legal coordination (M1-24)

Lead: PP, Participants: KL, CM, IR, IViR, KT, ToN, BnL, WMDE

The legal coordinator provides ongoing legal advice on the development of privacy policies, terms of service and other legal necessities. The legal coordinator monitors the intellectual property rights created during the project and ensures open dissemination of project results. This task includes creating a consortium agreement by M1, and ensure that it is signed by all partners by M3.

Milestones

➔ M6.1 All partners signed a consortium agreement (M3).

Deliverables

- ➔ D6.1 Data management plan (M6).
- ➔ D6.2 Interim report (M13).
- ➔ D6.3 Final report (M25).

List of Deliverables

Deliverable	Name	WP #	Lead	Type	Dissemination level	Month
D1.1	Online documentation of the Octopus data model research and design	1	PP	OTHER	PU	9
D1.2	Online documentation and initial implementation of the Octopus API research and design	1	PP	DEM	PU	10
D1.3	Uniform user interfaces available based on a corporate identity, user interface research and design (M11).	1	PP	DEM	PU	11
D1.4	Discovery features implemented on Project Octopus	1	PP	DEC	PU	14

D1.5	Metrics features implemented on Project Octopus	1	PP	DEC	PU	16
D1.6	Ingestion pilots finished from three different sources	1	CM	OTHER	PU	18
D1.7	Core infrastructure complete, stable, source available	1	PP	OTHER	PU	24
D2.1	Overview of suitable openly available algorithms to support the discoverability features of the project	2	KT	OTHER	PU	6
D2.2	Implementation and documentation of the algorithms implemented in the project	2	KT	OTHER	PU	12
D2.3	Report on data requirements and implementation of whitelisting of media files for public open reuse	2	KL	R	PU	4
D2.4	Processes defined and implemented for continuous refinement of data with autonomous agents	2	CM	OTHER	PU	22

D3.1	Research report on personal data protection issues related to open online repositories for provenance information	3	IViR	R	PU	M7
D3.2	Research report on Provenance information as a driver for more efficient licensing and rights clearance	3	IViR	R	PU	12
D3.3	Research report on the Legal validity of provenance information in the context of the copyright framework	3	IViR	R	PU	M20
D3.4	Governance model for Project Octopus	3	IViR	R	PU	M24
D4.1	Report of feedback from user groups	4	IR	R	PU	M10
D4.2	Online proceedings of symposium	4	IR	R	PU	M12
D4.3	User guidelines for use and participation in Project Octopus	4	IR	R	PU	M14
D4.4	Report on outreach	4	IR	R	PU	
D5.1	Strategic plan for communication and dissemination	5	KL	R	PU	M3
D5.2	The project website	5	KL	DEC	PU	M6

D6.1	Data management plan	6	KL	R	PU	M6
D6.2	Interim report	6	KL	R	PU	M13
D6.3	Final report	6	KL	R	PU	M25

3.2 Management structure and procedures

3.2.1 Organisational Structure

The project management deals with the project's organisational, administrative, financial and operational issues, as well as the decision-making processes. The Project Executive Board seats all Work Package leads. The coordinator decides about the distribution of the funding and approves financial and technical project reports to the European Commission. Work Package leads represent all partners within their Work Package.

The coordinator (KL) is in charge of all the coordination and management activities. Following tasks are performed by the coordinator:

- ➔ Chair the Project.
- ➔ Supervise the various work packages (WP), the objectives for the project including the quality and timelines of the various findings and reports to the EU.
- ➔ Manage the overall legal, contractual, financial and administrative of the consortium.
- ➔ Take the intermediary position in communication between the contractors and the Commission.
- ➔ Receive all payments made by the Commission to the consortium and administer the Community.
- ➔ Arrange all contributions regarding the allocation of resources between contractors and activities in accordance with the contract and the decisions taken by the consortium. The coordinator ensures that all the payments are made to contractors without unjustified delay.
- ➔ Inform the Commission of the distribution of the funds and the date of transfers to the contractors.
- ➔ Manage a database with relevant contacts, project documents and a file with administrative notes.
- ➔ Prepare meetings and minutes of the Project Executive Board.

The coordinator is supported by PP for legal coordination within the project and technical coordination (T6.3 and T6.4).

The coordinator endeavours to resolve any conflicts that arise at the lowest possible level. Every work package leader is responsible to solve minor issues within his/her work package. Only if this fails the conflict is discussed within the Executive Board. In case a partner is unable to fulfil its tasks, a (virtual) General Assembly of project partners is organised. It can lead to the decision to remove this partner from the project consortium. In case this partner is dismissed, or in case a partner should decide to drop out of the project, the General Assembly also decides democratically on the admission of a new partner to take over the tasks from the former partner.

The consortium described a well defined work programme and allocated tasks and responsibilities in full cooperation of all consortium partners beforehand. Nevertheless with the project's runtime of 24 months some expected and unexpected situations may occur, that may influence the successful outcome of the project. To reduce the overall risk the following steps have been taken:

- ➔ The establishment of a strong management structure.
- ➔ The coordinator with support of all consortium partners monitors the project continuously.
- ➔ Progress, achievements of milestones and task efforts to identify possible upcoming risks and problems immediately. This enables the undertaking to counter measures at an early stage.

3.2.2 Milestones

Milestone number	Milestone name	Related WPs	Estimated date (Month)	Means of verification
M1.1	Initial user interaction research complete	WP1	3	Written analysis.
M1.2	Initial data model, API designs and initial implementations; initial design and user interaction implementation	WP1	6	Code on GitHub, user specification released publicly.
M1.3	Database schema and API updated based on feedback from	WP1	9	Code on GitHub, API

	T1.3 implementation; user interface implementation uses API backend; API ready for beta users; initial content discovery and dashboard designs; research on external data models for ingestion complete			validated by a beta user
M1.4	Database schema, API, and user interface updated based on feedback from T1.4 and T1.5 implementations; initial content discovery and dashboard implementations; initial ingestion of provenance info from all pilots	WP1	12	Code on GitHub, API validated in beta
M1.5	Database schema and API updated based on feedback from T1.6 and WP2; ingestion from pilots complete	WP1	15	Information in database
M1.6	User interface, content discovery and dashboards updated based on feedback from T2.3 and T4.3	WP1	24	User interface tested
M2.1	Create a representative sample of media files	WP2	1	Representative sample available
M2.2	Perform initial assessment and validation of algorithms	WP2	3	Code on GitHub
M2.3	Implement, document and demonstrate algorithms for end-user use	WP2	6	Algorithms tested in beta
M2.4	Ascertain minimum viable information required of the data model for whitelisting	WP2	3	Written analysis

M2.5	Implement a whitelist in Outofcopyright.eu using data from the project	WP2	12	Code in GitHub
M2.6	Identify tasks suitable for autonomous agents (informed by T4.1)	WP2	12	Analysis
M2.7	Implement three autonomous agents and demonstrate against the project database	WP2	18	Analysis of the results of the demonstration
M3.1	Map personal data protection issues expected to arise during the implementation of Project Octopus	WP3	2	Written analysis
M3.2	Design principles to ensure compliance of Project Octopus with personal data protection rules	WP3	5	Written analysis
M3.3	Map potential interactions between open provenance systems and copyright licensing and rights clearance practices	WP3	8	Written analysis
M3.4	Formulate initial principles for open web based provenance systems	WP3	10	Written analysis
M3.5	Define provenance reliability principles and models	WP3	10	Written analysis
M4.1	Report on Creative Communities	WP4	6	Written report
M4.2	Report on Remix Communities	WP4	6	Written report

M4.3	Report on GLAMs	WP4	6	Written report
M4.4	Continuous documentation of feedback to development in online collaborative too	WP4	6, 10, 14, 18	Written analysis
M4.5	Symposium	WP4	10	The meeting
M5.1	Kick-off meeting	WP5	1	The meeting
M5.2	All partners meeting	WP5	14	The meeting
M5.3	A visual identity for Project Octopus	WP5	4	Visual representation
M6.1	All partners signed a consortium agreement	WP6	3	The agreement

3.2.3 Critical risks for implementation

Description of risk	**Work package(s) involved**	**Proposed risk-mitigation measures**
Too few algorithms available to implement in an open manner.	WP1, WP2	Partners CM and KT have already worked with algorithms so they know these.
Part of the user interface is discovery. If similarity recognition is not well-implemented, this might not be interested for end users.	WP1, WP2	Define other way to make discovery on the user interface interesting, by not letting it all depend on the media recognition techniques.

While Switzerland has stated that it will fund H2020 parties based in Switzerland it might not provide funding for KT.[24]	all	KT has a compartmentalised task in this project. Also CM is capable of taking over tasks from KT. The project can still achieve most of its goals without KT, however it would lose some of its excellent position.
No community is interested in provenance curation.	WP4	Large harvests of provenance information (T1.6) in combination with programmatic enhancements of provenance information (T2.3).
Substantial privacy issues in collection already publicly available provenance information.	all	Research into privacy issues (T3.1) to inform data model (T1.1) to mitigate possible privacy issues.
A consortium partner fails to deliver or ceases operations.	all	Regular monitoring of progress both on the WP level as well as on the consortium level.

3.3 Consortium

A total of 9 consortium partners is involved in this project. The consortium reflects a mix of partners that are vested in technological development, research, and community outreach. Most of the partners have been formerly engaged with Creative Commons, which informs their interest in the issue of provenance information.

3.3.1 Technological Excellence

The consortium has considerable technological excellence in the topic of provenance information and the creative and publishing sector.

PP hosts substantial technical skill. Their latest project contributoragreements.org shows their expertise in the intersection between legal matters and technological development. Their key personnel consists of founders of media platforms and a former CTO of Creative Commons.

[24] See http://www.sbfi.admin.ch/h2020/index.html?lang=en

KL is the architect of media-sharing platform openimages.eu, a platform that hosts openly licensed video files and Wikipedia's largest contributor of videos. It is the architect of OutOfCopyright.eu, a platform that helps you indicate whether intellectual property rights apply to a given creative work in a given jurisdiction of Europe. Finally, KL is the main architect of the Europeana Licensing Framework, a standard setting (technical) infrastructure that clearly communicates the reuse permission of Europe's cultural heritage.

Commons Machinery has already developed a proof of concept to provide provenance information online. With its project elog.io, it developed a plug-in to give back provenance information of works, based on image similarity recognition. It ingested most images of Wikimedia Commons, roughly 20 millions images.

Technology development partner KT has also worked on similarity recognition in images (maps) before and expands their existing algorithms. PP has been working on drawing connections between similar (cultural) works and images on the Internet for Project Octopus. It is designing a data model of works and maps cultural flows. The partners are complementary in WP1 and 2. They have all been working on related developments, but each have their own field of expertise.

3.3.2 Legal Excellence

In addition to their considerable technological expertise, PP brings legal expertise to the table, in cooperation with the largest comparative institute of information law in Europe: IViR. Together with legal think tank IR and think tank KL they research issues to back up legal aspects of the provenance system.

3.3.3 Community Excellence

Dissemination and communication to the wider public is taken care of by our community partners WMDE, IR, ToN and the BnL. These partners also offer content for the ingestion into the data model.

WMDE is the largest independant chapter of the Wikimedia Foundation. They support the activities of Wikimedia projects in Germany. They are also founder of Wikidata, the largest community on structured data. This includes provenance information. ToN is one of the largest communities of freely licensed music. It connects users to over 20.000 artists worldwide. It helps exposing these artists and offers the legal tools to a remix community to reuse creative works and pay for the works if that is required. The National Library of Luxembourg is an innovative leader in the GLAM sector. Their involvement in Europeana projects have contributed to renewal of the sector and correct licensing information for millions of works across Europe.

3.4 Resources to be committed

3.4.1 Summary of staff effort

	WP1	WP2	WP3	WP4	WP5	WP6	Total Person/ Months per Participant
Kennisland / KL	3	3	7	2	10	9	34
Peer Practice / PP	27	2	4	0	2	6	41
Commons Machinery / CM	10	17,5	3	0	0	3	33,5
iRights / IR	1	0	7	16	2	1	27
Instituut voor Informatierecht / IViR	0	0	17	1	0	1	19
Klokan Technologies / KT	4	10,5	0	0	0	1	15,5
Tribe of Noise / ToN	1	0	0	12	0	1	14
National Library of Luxembourg / BnL	1	1	1	6	0	1	10
Wikimedia Deutschland / WMDE	2	0	0	9	0	1	12
Total Person/ Months	**49**	**34**	**39**	**46**	**14**	**24**	**206**

4. Members of the consortium

4.1. Participants (applicants)

4.1.1. Kennisland (KL)

Kennisland (KL) is an Amsterdam based enterprising think tank with a public mission: to make society smarter, to empower people to learn, and to renew themselves continuously. KL develops solutions to questions that arise during the transition to a knowledge-driven society, and is part of the vanguard of that process. KL learns how this must be done by locating and supporting innovators, maximising knowledge development and knowledge sharing and translating expertise into practical interventions and innovation. KL shares the knowledge accumulated in doing so with as many people as possible, because knowledge only gains value when it is shared. Its approach ensures that top down and bottom up are no longer polar opposites but instead, reinforce one another. KL works with governments, businesses, knowledge institutes and social organisations that share our ambitions.

Key personnel

Tessa Askamp [F] is an advisor copyright and open source. Tessa works on technological projects Europeana Creative and Europeana Cloud. She mediates between policy recommendations and their requirements for development parties.

Paul Keller [M] is a director of KL. As a senior copyright policy advisor Paul initiates new projects, advises governments, cultural heritage institutions and other organisations on open approaches to copyright policy. He is public project lead for Creative Commons Netherlands, board member for Creative Commons Global, and a board member of Europeana.

Maarten Zeinstra [M] is an advisor copyright and technology. Maarten has expertise as technology broker in the field of copyright and heritage. He has translates legal concepts and policy requirements into technological products, particularly in the field of copyright law. His work includes designing and building the architecture of the OutOfCopyright.eu platform, acting as the requirements engineer for the Europeana Licensing framework and lead architect for the open video distribution platform Open Images.

Relevant publications, products, services, and projects

Relevant publications:

- Paul Keller, Jan Müller, Sandra den Hamer & Marens Engelhard. *Beelden van het Verleden - 7 jaar Beelden voor de Toekomst* (English: Images of the Past - 7 years Images for the Future). March 2015. Available at: http://beeldenvoordetoekomst.nl/publicatie/BVDT_eindpublicatie_web.pdf

- ➔ Paul Keller et al. *The Public Domain Manifesto.* Available at: http://www.publicdomainmanifesto.org/node/8

Relevant other products, websites and projects:

- ➔ Maarten Zeinstra and Tessa Askamp launched the online platform *OutOfCopyright.eu* as part of the Europeana Awareness project. Available at: http://outofcopyright.eu/
- ➔ Paul Keller and Maarten Zeinstra released metadata to the public domain in the Europeana Licensing Framework. For more information: http://pro.europeana.eu/web/guest/available-rights-statements
- ➔ Maarten Zeinstra developed the *Public Domain Charter*. This consists of decision trees for 26 jurisdictions that answer the question if a work is in the public domain or not. Available at: http://archive.outofcopyright.eu/
- ➔ Paul Keller and Maarten Zeinstra co-developed *Open Images*, an open media platform that gives access to a selection of archival materials that are open for creative reuse. Available at: http://www.openbeelden.nl/

4.1.2. Peer Practice (PP)

PP is a consulting firm specialised in the nexus of international intellectual property, information technology law, and cutting-edge technical development. It is a team with highly relevant experience in the field of (open) content licensing. Their approach is interdisciplinary and cross-cultural. Their projects work on the intersection between legal, management, policy, and technical strategies, drawing from academic theory and large-scale Internet service development.

Key personnel

Catharina Maracke [F] is the founder of PP. Her work and interests include copyright law and policy, standardisation efforts for public licensing schemes, and the general interaction between law and technology. She is also an associate professor at the Graduate School for Media and Governance, Shonan Fujisawa Campus, at Keio University and a fellow at the Berkman Center for Internet & Society at Harvard Law School. She is a member of the Global Agenda Council on the Intellectual Property System at the World Economic Forum and a member of the Defensive Patent License Advisory Board. As international Director at Creative Commons she has overseen the international license porting project for more than 3 years. She has also served as a board member for iCommons and the OpenCourseWare Consortium.

Mike Linksvayer [M] is the technical head of PP. He is a manager and technologist, who is passionate about building effective, scalable organisations, exploiting elegantly pragmatic technologies, and achieving huge increases in social welfare and consumer surplus. His specialities are in software management, semantic web, free/open source software.

Jon Phillips [M] is developer at PP. He has also founded the Fabricatorz and Openclipart, the largest community curated vector art repository. He is passionate about graphic design, open

source, social networking and entrepreneurship. In the past Jon has worked as the Community Director at Creative Commons.

Christopher Adams [M] is a designer and developer at PP. He is passionate about open source, and different designs to deliver products that are as appealing as possible. He also works for a variety of other organisations including Fabricatorz, Qi Hardware, AikiLab, and Sharism.org. Christopher has extensive experience building systems for the art, design, and publishing industries.

Relevant publications, products, services, and projects

- ➔ Catharina Maracke, Copyright Management for Open Collaborative Projects – Inbound Licensing Models for Open Innovation, 2013 SCRIPTed Volume 10 Issue 2, available at http://script-ed.org/?page_id=1025
- ➔ Catharina Maracke, Cultural flat rate, digital libraries, Creative Commons – What role for collecting societies in the 21st century? GRUR Int. 2010, 671 (Heft 8/9).
- ➔ Catharina Maracke, Intelligent Multimedia: Sharing Creative Works in a Digital World, ed. together with Daniele Bourcier, Melanie Dulong de Rosnay, and Pompeu Casanovas (European Press Academic Publishing in Florence, Italy 2010).
- ➔ Catharina Maracke, Creative Commons International – The International License Porting Project, JIPITEC Vol. 1, 2010, available at http://www.jipitec.eu/issues/jipitec-1-1-2010/2417
- ➔ Catharina Maracke & John Hendrik Weitzmann, Creative Commons – ein rechtliches Laienwerkzeug in der digitalen Welt, OEKOM, Berlin 2008.
- ➔ Catharina Maracke, A Delimitation of Design and Copyright, Institute of Intellectual Property, Tokyo 2006.
- ➔ Catharina Maracke, Die Entstehung des Urheberrechtsgesetzes von 1965, Duncker&Humblot, Berlin 2003.

4.1.3 Commons Machinery (CM)

CM is a research & development business, dedicated to putting digital works distributed online into their right context. In 2013-2015, it researched and developed a repository for provenance information, and implemented algorithms for searching the repository using perceptual hashes for images. It also led the implementation of browser plug-ins that facilitated this searching of information for images found online. All implementations developed are released as free and open source software for everyone to use. CM was founded in 2013 and is a member of the

International Press and Telecommunications Council (IPTC), which specialises in developing metadata standards for images.

Key personnel

Jonas Öberg [M] is the founder and CEO of CM. He leads the development team and is keen on all things open. Jonas has been the regional coordinator in Europe for Creative Commons and is currently also the executive director CEO of the Free Software Foundation Europe.

Artem Popov [M] is a software engineer who have been instrumental in building the Elog.io provenance catalog. He is a keen musician and have previously worked on open source projects for audio and graphics.

Peter Liljenberg [M] is a senior architect, having worked with large scale distributed databases and software. His expertise is in distributed and scalable software solutions and architectures for big data storage.

Relevant publications, products, services, and projects

- ➔ Elog.io - a distributed provenance catalog, available from htp://elog.io/
- ➔ Peter Liljenberg, Expressing rights with metadata: state of the art standards, 2013, available from http://bit.ly/1CC1DDH
- ➔ Haoqing Yinhe and Jing Liu, Exploring Challenges in Embedding Metadata of Licence Information in Digital Work: A Case Study with Experts and End Users, 2013, available from http://subs.emis.de/LNI/Proceedings/Proceedings220/3108.pdf
- ➔ Jonas Öberg, A Distributed Metadata Registry (Blog post), 2013, available from http://commonsmachinery.se/2013/07/a-distributed-metadata-registry/
- ➔ Jonas Öberg, Why we Need/Don't Need a Registry of Works (Blog post), 2014, available from http://commonsmachinery.se/2014/01/why-we-needdont-need-a-registry-of-works/

4.1.4 iRights.Lab (IR)

IR is a Berlin based independent think tank. It designs solutions for dealing with the challenges our societies face because of digitisation. Its team of experts in the fields of copyright and privacy law, journalism, licensing models, community outreach and other issues provides research, develops and implements strategies, and offers a context for interdisciplinary discussions – to stakeholders from the public sector, civil society, business, and politics. Its mission: to help use the opportunities of digitisation for the greatest possible benefit to society.

Its work includes developing and organising events revolving around current issues like the right to be forgotten, big data, Open Educational Resources and the future of copyright and creative industries.

IR has grown out of the iRights.info platform, Germany's most important news site on legal issues in the digital age. iRights.info was founded in 2004 and won - amongst other accolades - the Grimme Online Award, Germany's most prestigious prize for online journalism.

Key personnel

Valie Djordjevic [F] is an editor at IR and works as a lecturer and workshop trainer on digital culture, writing, social media and copyright at private and public institutions all around Germany. Valie was a member of Internationale Stadt Berlin, one of the first net culture projects in Germany, in 1996. Since then she worked in different art and cultural projects including mikro e.V., a Berlin-based association examining the different facets of media culture, the cyberfeminist group Old Boys Network or the Media Arts Lab at Künstlerhaus Bethanien. From 2009 to 2011 she was part of the team of IUWIS, a copyright infrastructure for research and education based at the Humboldt University Berlin. She is a co-moderator and administrator (together with Diana McCarty, Kathy Rae Huffman and Ushi Reiter) of the mailing list Faces, one of the first lists for women working with art and media. She writes fiction and essays on art and culture and occasionally even makes art herself.

Dr. Paul Klimpel [M] is an attorney with a background in the management of cultural institutions. He is specialised in the legal and technical dimensions of digital heritage. He holds a PhD from Humboldt University Berlin and was administrative director of Stiftung Deutsche Kinemathek, Germany's premier film museum with a collection of 13.000 works, from 2006 to 2011. He also served as general manager of the German network of media libraries before joining IR as head of the culture lab. He serves as coordinator for cultural heritage of the Internet & Gesellschaft Collaboratory and publishes widely on digitisation in GLAMs. Since 2010 he organises the annual international conference ‘Zugang Gestalten’ (‘Shaping Access’) which has developed into the essential event on the topic in Germany, convening 200 experts from all over the world.

Susanne Lang [F] is a psychologist by training. She has specialised on financial and organisational project management as well as community outreach. She was a core member of the information platform ‘Verbraucher sicher online’ (‘Consumers staying safe online’) funded by the German Ministry for Consumer Protection, where she wrote manuals and how-to articles, advising consumers on ways to safeguard their privacy and securely conduct all kinds of communication online. She is a certified coach and conducts trainings on campaigning, change management and project management for NGOs in Germany and internationally.

Matthias Spielkamp [M] is co-founder and managing editor of the online magazine iRights.info and co-founder of iRights.Lab. As a journalist, Matthias is a frequent commentator on the consequences of digitisation on society for national German newspapers, magazines, online media and radio stations. He has worked with journalists in Germany, South-Eastern Europe, Asia and the Middle East for institutions like Deutsche Welle, the International Institute for Journalism of GIZ and others for 15 years. Matthias is frequently invited to speak at conferences like re:publica or the Global Media Forum and conceived and organised the Young Media Summits 2010 and 2011 in Cairo, Egypt. He testified before three committees of the German Bundestag on future developments of journalism, online journalism and copyright

regulation. Matthias has co-authored three books on journalism and copyright regulation and holds master's degrees in Journalism from the University of Colorado at Boulder and Philosophy from the Free University of Berlin.

John H. Weitzmann [M] is accredited as an attorney in Germany. He is partner at iRights.Law in Berlin, a law firm specialising in Open Licensing practices, cultural heritage consulting, legal aspects of science and education as well as data protection compliance, and supports iRights.Lab as a consultant. Since 2006 he has been representing Creative Commons Germany in the capacity of legal project lead (pro bono) and, inter alia, initiated round table debates between established stakeholders and entrepreneurs on collective management of rights. John is also founding member of the digital rights group Digitale Gesellschaft e.V. He publishes articles on topics of law in the digital world on a regular basis and supports the editorial work of the online magazine iRights.info providing legal expertise.

Relevant publications, products, services, and projects

- ➔ International Conference Zugang gestalten (Shaping access), annually since 2010, approx. 200 international participants, Berlin (http://www.zugang-gestalten.de/shaping-access-more-responsibility-for-cultural-heritage/).
- ➔ E. Euler & P. Klimpel (eds.), Der Vergangenheit eine Zukunft - Kulturelles Erbe in der digitalen Welt (A future for the past - cultural heritage in a digital age), Berlin: iRights.Media, 2015.
- ➔ J.H. Weitzmann, Vielfalt für die Ewigkeit. Was Creative Commons für alle Gedächtnisinstitutionen so interessant macht (Diversity forever - Why Creative Commons appeals to heritage institutions), in E. Euler & P. Klimpel (eds.), Der Vergangenheit eine Zukunft - Kulturelles Erbe in der digitalen Welt, Berlin: iRights.Media, 2015, ISBN 978-3-944362-06-9.
- ➔ V. Djordjevic, L. Dobusch (eds.), Generation Remix, Berlin: iRights.Media, 2014, ISBN 978-3-944362-02-1.
- ➔ V. Djordjevic, M. Spielkamp et al., Kopieren, Bearbeiten, Selbermachen – Urheberrecht im Alltag (Copy, Remix, Create – a Copyright Guide for Every Day, with Djordjevic et al.), Bonn: Bundeszentrale für politische Bildung, 2008.

4.1.5 The Institute for Information Law (IViR)

The IViR is part of the Faculty of Law of the University of Amsterdam. The Institute is the largest research facility in the field of information law in Europe and probably in the world. During a national evaluation of legal research programmes, the institute received the highest grades of all programmes in its field of expertise. The Institute employs approximately 25 qualified researchers who actively study and report on a wide range of topics within the domain of information law. Due to its large scale and its expert multinational and multilingual research

staff (working languages include English, German, Dutch, French, Italian, Spanish, and Portuguese), IViR is able to gather and analyse much of the required information regarding the implementation and application of the relevant EU and national legal sources. An important asset of the Institute is its vast network of expert correspondents across the European Union, mostly professors of copyright law or specialised lawyers, that can assist in providing up-to-date and reliable information on legislative initiatives, and any relevant ensuing case law. The Institute has at its disposal a well equipped Documentation Centre containing a wealth of information resources in the field of European copyright and media law (books, periodicals, databases, reports, legislation and legislative proposals, grey literature, etc.).

IViR and its international staff regularly give advice to the European Commission, the European Parliament, the Council of Europe, WIPO and national governments on matters related to information law and policy. At the European level, IViR is part of the multidisciplinary network of partners of the European Audiovisual Observatory, the information network founded in 1992 by 33 European states and the European Commission to serve the audiovisual industry. Through personal networks and formal and informal arrangements, IViR researchers have direct access to scientific fields ancillary to information law, such as information market economics and communications sciences. IViR's research program for 2012 – 2016 can be found at http://www.ivir.nl/research/IviR_Research_Program_2012_2016.pdf .

The institute has a vast experience in the organisation of conferences and symposia, and conducts research into the fundamental rights that give shape to the information process (freedom of communication, privacy and intellectual property) and into the rules of the information market (media law, telecommunications law, competition law, consumer protection).

Key personnel

Dr. Lucie Guibault, [F], associate professor, is specialised in international and comparative copyright and intellectual property law. Lucie Guibault has been carrying out research for the European Commission, Dutch ministries, UNESCO and the Council of Europe. Her main areas of interest include copyright and related rights in the information society, open content licensing, collective rights management, limitations and exceptions in copyright, and author's contract law. She has been involved as legal partner in Creative Commons Netherlands since 2005 and in projects related to Europeana (EuropeanaConnect and Europeana Awareness) since 2009.

Dr. Stef van Gompel [M], is specialised in intellectual property law and, in particular, in national and international copyright law. He has written various articles and book chapters on this topic. He is secretary of the Dutch Copyright Committee that advises the Minister of Justice of the Netherlands on copyright-related matters. He is also a member of the editorial board of the Dutch copyright journal *AMI* (*Tijdschrift voor Auteurs-, Media- & Informatierecht*) and chairman of the Study group on the history of copyright of the Dutch copyright organisation, *Vereniging voor Auteursrecht* (VvA). At IViR, Stef is currently working as a postdoc researcher, preparing the contribution of the Netherlands to the *Primary Sources on Copyright (1450-1900)* project,

edited by Lionel Bently (University of Cambridge) and Martin Kretschmer (University of Glasgow).

Dr. Simone Schroff [F] - is a researcher in copyright law at the IViR. She completed a BA in History and Politics at Keele University and a MA European Governance at Exeter University. She has since gained a PhD from the University of East Anglia (UK) (distinction) where she defended her thesis *The Evolution of Copyright Policy 1880-2010: A Comparison between Germany, the UK, the US and the International Level*. She specialises in the qualitative, quantitative and comparative analysis of copyright law and policy. Her main areas of interest are copyright and related rights in the digital context, the driving forces of copyright development and the framing of copyright policy. Simone Schroff has published parts of her work and was awarded the *Outstanding Publication by a Postgraduate Research Student Award* from the University of East Anglia Law School for her article titled *'The (Non) Convergence of Copyright Policy'*.

Relevant publications, products, services, and projects

- ➔ 'Orphan Works Directive', in T. Dreier and P.B. Hugenholtz (ed.), Concise European Copyright Law, Alphen aan den Rijn, Kluwer Law International, forthcoming 2015.
- ➔ L. Guibault, 'Collective Rights Management Directive', in I. Stamatoudi & P. Torremans (eds.), EU Copyright Law, Cheltenham: Edward Elgar, 2014 ISBN978 1 78195 242 9.
- ➔ L. Guibault, Copyright Limitations and Contracts, An Analysis of the Contractual Overridability of Limitations on Copyright, The Hague: Kluwer Law International, February 2002, 392 pp., ISBN/ISSN 9041198679.
- ➔ L. Guibault & C.J. Angelopoulos (eds.), Open Content Licensing: From Theory to Practice, Amsterdam: Amsterdam University Press, 2011.
- ➔ L. Guibault & P.B. Hugenholtz (eds.), The Future of Public Domain: Identifying the Commons in Information Law, The Hague: Kluwer Law International, 2006, ISBN/ISSN 9041124357.

The institute has worked on multiple relevant projects funded by different local and international governments.

- ➔ EuropeanaConnect (2009 – 2011), funded by the European Commission (eContent+).
- ➔ OpenAire+ (2012-2014), funded by the European Commission (7th Framework Program).
- ➔ Creative Commons Nederland (2007 – 2015), funded by the Netherlands Ministry of Education, Culture and Science.
- ➔ Europeana Awareness (2012 – 2014), funded by the European Commission (CIP-ICT-PSP-2011-5).
- ➔ Study on the remuneration of authors and performers for the use of their works and the fixations of their performances MARKT/2013/080/D (2014 – 2015), funded by the DG Internal Market.

➔ Study on extended collective licensing as a solution for mass-digitisation of collections of cultural heritage institutions (2014), funded by the Netherlands Ministry of Education, Culture and Science.

4.1.6 Klokan Technologies (KT)

KT is a Swiss company specialised in raster image processing, online map publishing, geographic information retrieval and applications of open-source software for the cultural heritage sector. KT is also the developer of MapTiler software, which is a popular product for user-friendly web 2.0 map publishing in a form of overlays for Google Earth and Google Maps. The software is used by NOAA, NASA, Apple, Google, US Forest Service, European Commission Joint Research Center and several commercial companies, universities, libraries and government organisations all over the world.

The company is contributor to various open-source software projects including IIPImage JPEG2000 IIIF, Omeka, OpenLayers, WebGL Earth, OL3-Cesium and GDAL.

The team members actively participate in research projects with cultural heritage institutions. KT team developed a crowdsourcing platform for assigning precise geolocation to the scanned maps called Georeferencer and also an online search engine for old maps called OldMapsOnline.org, which allows comfortable search in almost half of a million out-of-copyright maps from the world prominent map libraries and private map collections, including the British Library, New York Public Library, Harvard Library, David Rumsey, USGS, NOAA, and several university libraries, agencies, national libraries and archives from Europe, America and Australia. Indexed are maps from 16th century to early 20th century of various scales.

Thanks to the expertise in the large scale raster image processing, experience from development of a unique search engine with a specific indexing mechanism (MapRank®), access to world largest collection of online old maps and thanks to ongoing practical research in the field of image similarity.

Participant from a non-EU member state

KT is based in Switzerland, they are the only participant from a non-EU member state. The organisation brings expertise in software development in similarity recognition of digital images. Due to its expertise and experience with Project Octopus' user groups, KT is an excellent candidate to provide the required services for this project.

Key personnel

Petr Pridal [M] is a managing director and technical leader of the company. He gained his Ph.D. in the field of cartography and geodesy and a master degree in computer science. For over a decade he acts as a consultant, programmer and entrepreneur. As founder of Klokan Technologies GmbH, he drives his company and a small team of programmers to be the innovative developer of mapping applications empowering people, companies and institutions

to search, publish and enjoy the real value of maps they own. Petr participated on student research programmes at Google. He also created several popular open-source projects, including a software which was downloaded by more than 7 million people.

Vaclav Klusak [M] works as a developer and geospatial consultant at KT. He is specialised on high-performance database applications and cloud development. His analytical expertise is backed by knowledge of several programming languages (C/C++/Python/Java). In the past he participated on Google Summer of Code project in geospatial domain. Vaclav has a master's degree in computer science.

Relevant publications, products, services, and projects

- ➔ Southall, H., Pridal, P. (2012): Old Maps Online: Enabling global access to historical mapping, e- Perimetron, vol. 7, issue 2.
- ➔ Fleet, Ch., Kowal, K., Pridal, P. (2012): Georeferencer: Crowdsourced georeferencing for map library collections, D-lib Magazine, vol 18, issue 11.
- ➔ Oehrli, M, Pridal, P., Zollinger, S., Siber, R. MapRank (2011): Geographical search for cartographic materials in libraries, D-lib Magazine, vol 17, issue 9/10.
- ➔ MapTiler (Software) http://www.maptiler.com/
- ➔ MapRank Search (Software) and OldMapsOnline.org search engine.

4.1.7 Tribe of Noise (ToN)

ToN is an online community that connects musicians with media professionals and businesses around the world in need for high-quality and all rights included music. Individual artists preserve their music rights and at the same time take advantage of the collective business deals, exploitation models and contacts, facilitated by ToN.

ToN represents over 25.000 artists from 185 countries and supplies music (licenses) for film, TV, video production, gaming and in-store media industry.

Key personnel

Hessel van Oorschot [M] is senior advisor and experienced entrepreneur with background in online business models, digital media and building international teams to execute disruptive ideas. Co-author of four successful workbooks (commissioned by the Dutch Ministry of Economic Affairs) enabling business owners to grow their online business.

Els Bond [F] is experienced in customer relations and online community management. Bachelor of Economics in International Music Management. Operationally involved in the Tribe of Noise rights holders community and the development of tools, procedures and mechanisms to improve the engagement with rights holders worldwide.

Relevant publications, products, services, and projects

- ➔ The following publications were commissioned by the Dutch government of Economic Affairs, Syntens, the Chamber of Commerce and/or trade associations MKB Nederland & VNO-NCW. In addition, we developed and executed eLearning modules, train-the-trainer sessions and workshops. Total number of (creative industry) entrepreneurs reached > 150,000.
    - ◆ 2003: Workbook “WWW voor het MKB” | WWW for SMEs (http://pundamilia.zinnebeeld.nl/downloads/www_voor_het_mkb.pdf).
    - ◆ 2006: Workbook “Klaar voor Digitaal ondernemen?” | Ready for eBusiness? (http://pundamilia.zinnebeeld.nl/downloads/klaar_voor_digitaal_ondernemen.pdf).
    - ◆ 2009: Workbook “Op weg naar slimmer samenwerken” | The journey to smart partnerships (http://pundamilia.zinnebeeld.nl/downloads/op_weg_naar_slimmer_samenwerken.pdf).
    - ◆ 2010: Workbook “Ondernemen met Sociale Netwerken” | Social Networks for business owners (http://www.pundamilia.nl/downloads/ondernemen_met_sociale_netwerken.pdf)
- ➔ Design, development and execution of a global community (2008 – today) for rights holders using Creative Commons as the main structure for the legal framework. Resulting in over 25,000 participating rights holders from 185 countries. https://en.wikipedia.org/wiki/Tribe_of_Noise
- ➔ Development of professional user groups in the creative industry like NewMusicIndustry.org with 8,440 members today.
- ➔ 2010: Development of a global creative contract, the “Non Exclusive Exploitation Contract”, as a spin-off of Creative Commons 3.0 By – Share Alike. Over 15,000 items globally licensed under a NEEC today.
- ➔ 2006 – 2010; Development and execution of international trade missions to creative industry events like the Game Developers Conference in San Francisco, Vancouver Digital Week, Leipzig Games convention, Midem (France) and SXSW in Austin, Texas.

4.1.8 National Library of Luxembourg (BnL)

Its role and status as the national library of Luxembourg are defined by the law of 25th June 2004.[25] Its roles include:

- ➔ the collection and preservation of all printed national heritage (incl. digital born)
- ➔ building an encyclopedic international collection
- ➔ giving the best and widest possible access to it, including to the general public
- ➔ the management of the national union catalogue and library system[26]
- ➔ the management of the digital scientific library and its licensing consortium[27].

All collected works, including the digitised collection from www.eluxemburgensia.lu are availbale under a single search engine at www.a-z.lu.

Key personnel

Patrick Peiffer [M] has a Master in Library and Information Science from Humboldt University. He has more than ten years of experience in digital library services. Relevant projects include: designing and managing the national digital library consortium with its licensing and access services, rights clearance for the national library's digitisation projects and partner in both Europeana Connect and Awareness projects with a focus on licensing issues and services. Patrick is a member of the European Commission's 'Member State Expert Group on Digitisation' [28] since inception.

Guillaume Rischard [M] has a Computer Science BSc from the University of Luxembourg and also followed Intellectual Property classes. He has worked as a freelance consultant and developer and has experience in open licensing and community work through involvement in Open Street Map and Codeclub Luxembourgh projects. Currently, Guillaume is working for the National Library on several projects centered around metadata collection, transformation, enrichment and quality control. He supports Patrick Peiffer on IT and IPR related matters.

Relevant publications:

- ➔ "Libraries and publications of the future", 2013, Consortium Luxembourg, http://blog.findit.lu/wp-content/uploads/2013/04/Libraries-and-publications-of-the-future-online-new1.pdf

Projects:

- ➔ Patrick Peiffer was work package leader for IPR in Europeana Connect (2009-2011, eContentPlus) which developed the Europeana Licensing Framework, leading to, among other significant policy documents, to the largest ever addition of cultural heritage

[25] http://www.legilux.public.lu/leg/a/archives/2004/0120/2004A17981.html
[26] http://www.bibnet.lu
[27] http://blog.findit.lu/about
[28] http://ec.europa.eu/digital-agenda/en/member-states-expert-group-digitisation-digital-preservation

metadata under the CC0 metadata dedication.
http://www.europeanaconnect.eu/results-and-resources.php?page=8

- ➔ Patrick Peiffer continued active participation in the IPR workpackage under Europeana Awareness (2012-2014, FP7) and among others ran a series of expert workshops on Extended Collective Licensing,
http://www.bnl.public.lu/fr/actualites/communiques/2014/11/Europeana-Workshop/index.html
- ➔ Patrick Peiffer is a member of the current MSEG 'Member State Expert Group on Digitisation' Task Force on Sustainability of Europeana, participating in key discussions about value, impact and investment in digital cultural heritage.

Products:

- ➔ BNL is hosting the redeveloped (in 2014) Public Domain Calculator (www.outofcopyright.eu) as well as the (under development) rightsstatements.org project, both in Tier IV certified datacenters.

4.1.9 Wikimedia Deutschland (WMDE)

WMDE (Wikimedia Germany - Society for the Promotion of Free Knowledge) is an independent, charitable membership-based non-profit organisation serving as a Wikimedia chapter. According to the charter, the purpose of the organisation is to support the creation, collection and distribution of Free and Open Content to support equal opportunity and access to knowledge and education. WMDE was founded in 2004 as the first chapter of the global Wikimedia movement. Today, the organisation has 70 employees, 20,000 members, while over 6000 volunteer editors and photographers are active in Wikimedia projects.

WMDE supports and promotes these Wikimedia projects in Germany. These include the German-language Wikipedia, Wikimedia Commons (the database of 25 million media files), Wikidata and a number of projects that connect Wikimedia volunteers with GLAM institutions. WMDE works with and supports a large and diverse remix community.

Key personnel

Sebastian Sooth [M] is the Head of Volunteer and Idea Support Services at WMDE. He has developed Wikimedia's participatory support systems that enable volunteers to independently move their projects from idea to shared learning. He has almost two decades of experience in project management and consulting in the IT and digital knowledge sectors.

Dipl. Pol. Nicola Zeuner [F], manages partnerships and externally funded projects at WMDE. She has more than 20 years of experience in coordinating multi-sector coalitions and projects.

Publications:

- ➔ Buchem/Kloppenburg/Weichert: Charting Diversity. Working together towards diversity in Wikipedia, Berlin 2014.
https://commons.wikimedia.org/wiki/File:Charting_Diversity.pdf

- ➔ Vrandecic/Kroetzsch: Wikidata: A Free Collaborative Knowledge Base, New York 2014http://korrekt.org/papers/Wikidata-CACM-2014.pdf
- ➔ Kreutzer: Open Content - A Practical Guide to Using Creative Commons Licences, Berlin 2014.

Products:

- ➔ Wikimedia projects: https://en.wikipedia.org/wiki/Wikimedia_project
- ➔ Wikidata: free linked database and central storage for structured data https://www.wikidata.org/wiki/Wikidata:Main_Page
- ➔ Diversity-Aware Ranking Service http://ranking.render-project.eu/
- ➔ Toolkit for Knowledge Diversity in Wikipedia: https://tools.wmflabs.org/render/toolkit/

4.2. Use of third party resources

The Project Octopus consortium has expertise in technological development, legal research, community building and has a strong organisational coordinator. It develops all of the work described in its description of work. We do request the use of expert third party resources as part of one task.

4.2.1 Kennisland

No third parties involved.

4.2.2 Peer Practice

No third parties involved.

4.2.3 Commons Machinery

No third parties involved.

4.2.4 iRights

No third parties involved.

4.2.5 IViR

No third parties involved.

4.2.6 Klokan Tech

No third parties involved.

4.2.7 Tribe of Noise

No third parties involved.

4.2.8 National Library of Luxembourg

Does the participant plan to subcontract certain tasks	**YES**
T2.2 (Whitelist of Media Files: researches and implements methodologies to determine the rights status of creative works) is designed to integrate third-party service OutOfCopyright.eu. This platform has been developed by a third party software development company (NeoFacto) which given their existing involvement in the platform will be the most cost efficient developer for the extension foreseen as part of project octopus. The scope and size of involvement of the developers behind that project will not exceed 1.5 Person Months. Overhead created by structurally involving that party will not be efficient spending. We reserve €10.000 for adjustment of that platform.	
Does the participant envisage that part of its work is performed by linked third parties	**NO**
Does the participant envisage the use of contributions in kind provided by third parties (Articles 11 and 12 of the General Model Grant Agreement)	**NO**

4.2.9 Wikimedia Deutschland

No third parties involved.

5. Ethics and Security

Project Octopus has few relevant ethical and security considerations. The project collects provenance information from across the web which may cause a mosaic effect leading to privacy issues. The project's open ingestion and open publishing policy causes the possibility of importing and exporting data from/to parties outside of the EU. The project has no relevant security issues.

5.1 Ethics

Does this research involve further processing of previously collected personal data?

Project Octopus collects provenance information of creative works that are published online. It ingests provenance information from content platforms and aggregators. This provenance information can include the full name or pseudonyms of creators and rights holders and location where a single media file is used on the web. Potentially, a combination of data from a variety of sources can return sensitive personal data (mosaic effect).

While collecting information about the links between works and their creators/owners has a number of highly desirable consequences that form the basis of this proposal, it also means that such a system inevitably deals with the collection and processing of personally identifiable information. Task 3.1 is dedicated to research into privacy issues when collecting provenance information to ensure that no violation of privacy results from the project. This research informs the development of a data model used in project Octopus. This ensures that the combination of sensitive information is avoided.

Task 3.2 researches the relation between the copyright framework and provenance information. This informs the project's ingestion pilots (task 1.6) on whether they can structurally harvest provenance information from chosen sources. This informs governance principles of provenance systems that are produced as part of task 3.3. Finally, task 6.4 provides dedicated legal coordination to avoid raising privacy issues in this project.

Do you plan to import any material from non-EU countries into the EU?

As a consequence of the large aggregation of provenance data, data import and export between EU and non-EU countries is part of the project. Data from all over the Internet might become ingested into the project's database. In the initial stages of the project, data from Wikimedia Commons is used. This might contain data from anywhere on the web, including

from non-EU countries. Close examination of the ethical aspects of the use of this data is an integral part of the project.

Do you plan to export any material from the EU to non-EU countries?

Explicit export to non-EU countries is not part of Project Octopus. However, due to an open publication policy, all products of Project Octopus are available for reuse. This includes the database, algorithms, and all other source code that is accessible through a user interface, API and repository. It contains provenance information from citizens of EU-countries as well as from non-EU countries. Project Octopus makes provenance information widely available on the web, openly accessible to everyone.

5.2 Security

Activities or results raising security issues	**NO**
'EU-classified information' as background or results	**NO**