SEVENTH FRAMEWORK PROGRAMME
FP7-ICT-2009-6

BlogForever
Grant agreement no.: 269963

BlogForever: D3.3 Development of the Digital Rights Management Policy

Editor:	**Yunhyong Kim, Tracie Farell, Seamus Ross**
Revision:	First
Dissemination Level:	Public
Author(s):	Tracie Farrell (Populis), Yunhyong Kim (UG), Ed Pinsent (UL), Stella Kopidaki (Phaistos), Morten Rynning (CW), Ioannis Manolopoulos (AUTH), Olympia Papadopoulou (AUTH), Stratos Arampatzis (Tero), Ilias Trochidis (Tero), Despoina Zioga (Tero)
Due date of deliverable:	31/08/2013
Actual submission date:	31/08/2013
Start date of project:	01 March 2011
Duration:	30 months
Lead Beneficiary name:	UG

Abstract: This report presents a set of recommended practices and approaches that a future BlogForever repository can use to develop a digital rights management policy. The report outlines core legal aspects of digital rights that might need consideration in developing policies, and what the challenges are, in particular, in relation to web archives and blog archives. These issues are discussed in the context of the digital information life cycle and steps that might be taken within the workflow of the BlogForever platform to facilitate the gathering and management of digital rights information. Further, the reports on interviews with experts in the field highlight current perspectives on rights management and provide empirical support for the recommendations that have been put forward.

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)

The **BlogForever** Consortium consists of:

Aristotle University of Thessaloniki (AUTH)	Greece
European Organization for Nuclear Research (CERN)	Switzerland
University of Glasgow (UG)	UK
The University of Warwick (UW)	UK
University of London (UL)	UK
Technische Universitat Berlin (TUB)	Germany
Cyberwatcher	Norway
SRDC Yazilim Arastrirma ve Gelistrirme ve Danismanlik Ticaret Limited Sirketi (SRDC)	Turkey
Tero Ltd (Tero)	Greece
Mokono GMBH	Germany
Phaistos SA (Phaistos)	Greece
Altec Software Development S.A. (Altec)	Greece

History

Version	*Date*	*Modification reason*	*Modified by*
0.9	31/07/2013	Extracting and reformatting material authored by Tracie Farrell, Stella Kopidaki, Ed Pinsent, Vangelis Banos, Ilias Trochidis, Dimitrios Tektonidis, Morten Rynning. This was from the document on the Google drive.	Yunhyong Kim
0.99	29/08/2013	Taking into consideration all new contributions from Tracie, Stella, Ed, and Dimitrios, and feedback from Vangelis.	Yunhyong Kim
1.0	30/08/2013	Final typo and formatting corrections.	Ed Pinsent

Table of Contents

Executive Summary

This report details the results of BlogForever WP3 Task 3.3 Development of Digital Rights Policy described in the Project Description of Work (DoW), according to which: "the main objectives of this task are to develop a Digital Rights Management Policy (DRM) that will clearly define the access level and type of allowed use of all items stored in the BLOGFOREVER digital repository by different types of users. This task will include a survey of existing Web Archiving legal issues. Finally, this task will gather the surveys' results, identifying problems and solutions, including them into the BLOGFOREVER approach and developing the BLOGFOREVER License articulation (as described in WP3 description in section B1.3.1.1 of Part B)."

In the current context, BlogForever does not have an existing collection, which inhibits specific decisions with respect to these directions as there is no selected content nor intended users. The establishment of rights policies is heavily dependent on the content's context of creation, corresponding publishers, and the activities of the community that will be using and managing the collection (e.g. conditions of reuse in the learning context may differ from other contexts[1]). Even when these are specified, the laws are not definite and have not caught up to accommodate how we relate to digital information, especially with respect to information on the web and blog.

The current report is constructed to aid future curators of digital materials from the web and, more specifically, from blogs to establish their own digital rights policy to manage risks involved with various rights issues arising in the context of running a repository containing digital materials from the web. In particular, the report aims to aid the reader in answering four questions:

1. What issues exist for addressing digital rights management with respect to collections of web content and, in particular, blog content, and what policies have already been developed, on an institutional, national, and international level to address the issues?
2. What rights management capabilities might the BlogForever repository be able to provide (e.g. with respect to identifying rights management opportunities at key points in the digital information life cycle, authentication and authorisation technologies, and metadata assignment capabilities)?
3. Is there a common conversation among experts about rights management that might provide insight to groups involved in blog content management and supporting technologies?
4. What approaches for rights management might be developed in the future?

In relation to these, we explore a range of issues with a focus on digital rights with respect to information found on the web and in blogs (Section 2), revisit repository functionalities and examine opportunities within these workflows and survey how other archives have approached the task (Section 3), clarify the capabilities of the BlogForever repository and briefly analyse the potential for automated extraction of rights metadata (Section 4), discuss guidelines for cataloguing metadata (Section 5) and speak with experts in the field in the form of selected interviews to examine the ongoing conversation that might affect digital rights management for archives, library and repositories (Section 6). In Section 7, we summarise the finding from each section and conclude by making a few observations about digital rights management policy development for the future.

The report is intended to highlight the most immediate concerns to be addressed. It is not meant to be an exhaustive investigation. It is even doubtful that an exhaustive investigation is possible, because of the changing nature of the legal landscape, especially in relation to digital materials created on the web. The legislation related to this type of information is struggling to catch up with the way we interact on social media, that is, it is expected to go through many changes in coming years. Rights management will increasingly become a question of risk management, rather than a

1 http://www.reusablelearning.org/

question of protection measures. The general recommendation we are making here is to support common sense rather than legal sense. It should be noted that the authors of this deliverable are not lawyers. They are not qualified to give legal advice. The report is intended as a guideline only.

1 Introduction

This report describes the final results of Task 3.3 "Development of Digital Rights Management Policy" of the BlogForever project (EC FP7 Grant no. 269963).

According to the description of BlogForever WP3 Task 3.3 in the project Description of Work (DoW): "the main objectives of this task are to develop a Digital Rights Management Policy (DRM) that will clearly define the access level and type of allowed use of all items stored in the BLOGFOREVER digital repository by different types of users. This task will include a survey of existing Web Archiving legal issues. Finally, this task will gather the surveys' results, identifying problems and solutions, including them into the BLOGFOREVER approach and developing the BLOGFOREVER License articulation (as described in WP3 description in section B1.3.1.1 of Part B)."

The project aims to promote BlogForever adoption all around the world. Our intention, according to the DoW, was to create specific, written policy statements on access, licensing and other legal issues related to weblog preservation. However, we recognise that Copyright laws – and indeed many other laws associated with protecting rights – are not identical in every country, nor applied in the same ways.

In fact, in the current context, BlogForever does not have an existing collection, which inhibits specific decisions with respect to these directions as there is no selected content nor intended users. The establishment of rights policies is heavily dependent on the content's context of creation, corresponding publishers, and the activities of the community that will be using and managing the collection (e.g. conditions of reuse in the learning context may differ from other contexts[2]). Even when these are specified, the laws are not definite and have not caught up to accommodate how we relate to digital information, especially with respect to information on the web and blogs, as we will show in our discussions in Section 2.

This report aims to describe the recommendations of the BlogForever project for developing a set of rights management policies that acknowledge these difficulties and facilitate the ways in which users engage with the BlogForever platform, while maintaining a risk management strategy to stay within the law and protecting the rights of content owners.

1.1 Why Digital Rights Management?

This deliverable is about "rights management" in the very broad sense that it intends to enable users of the BlogForever platform to collect blog content, preserve it, and allow access to it, without infringing copyright, IPR, privacy or other legal issues. "Digital Rights Management" is still a contested term, but is often used in a commercial context to refer to mechanisms built into digital objects that control, restrict or deny copying of content, in order to protect copyright. It can also refer to the authoring tools used to create such mechanisms, often referred to as Technical Protection Measures[3].

Some have made the observation that DRM has largely consisted of restriction management, rather than the protection of rights[4]. These observations put forward the argument that "DRM creates a damaged good"[5]. While these observations were not introduced in the context of digital preservation, there is some validity in the statement in that mechanisms for restriction imposed on

2 http://www.reusablelearning.org/

3 Article 6, Copyright Directive, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001L0029:EN:HTML

4 http://www.defectivebydesign.org/what_is_drm_digital_restrictions_management

5 Ibid.

information through the use of technology (e.g. password protection; encryption; remote deletion) do give rise to new risks of information loss/inaccessibility, and, consequently, may pose a threat to digital preservation. Rights management policies will, further, determine the actions that can be taken within the repository as part of the digital preservation process (Coyle 2006). In the light of these issues, assuming that a digital repository manager is planning to support digital preservation, it is essential that an approach to rights management is developed that both serves to protect the rights of content providers, as well as, to support digital preservation.

To develop a digital rights management policy for a digital repository, we must start by defining the scope of rights we need to protect with respect to content, defining distribution and acquisition policies in relation to these rights, devising a plan of how you might enforce policies, and track content usage to ensure effectiveness of the policies (Collier, Piccariello & Robson 2004). Collier, et al. takes the approach that policies might include: assigning a license to content use (e.g. Creative Commons[6] and General Public License[7]), specifying permissions and requirements with respect to attribution, access, distribution, copying, and modification, selecting rights expression languages (e.g. Creative Commons Rights Expression Language[8] and Open Policy Language for the Digital Commons[9]) to promote the persistence of the rights information should the content change hands, identifying legitimate methods (e.g. Sharable Content Object Reference Model[10] in an agreed environment of learning objects) for tracking usage, making decisions concerning encryption and authentication, and determining the viability of using global persistent identifiers and/or handles (e.g. a general registry of digital objects such as Digital Object Identifier and special registries such as the ADL[11] learning objects registry) and rights information registries (e.g. Registered Commons[12] and Safecreative[13] for works with a Creative Commons License; Rights Metadata for Open archiving (RoMEO)[14] for publisher copyright information). This strategy is largely focused on issues surrounding copyright and related intellectual property rights.

As we will emphasise in later parts of this report, we would like to situate copyright in the context of wider concerns[15] of intellectual property such as trademark, patent and design, which may, in fact, easily arise within the context of web archiving. We would also like to examine changes that are occurring in relation to concepts of privacy, data protection, defamation, and illegal content, boundaries which are easily blurred and crossed when dealing with social media content where users upload, generate, embed, and interact with data freely (see further discussion in section 2).

To make the correct decisions regarding these matters, rights management policies must be developed to answer the following types of questions:

1. What issues exist for addressing digital rights management with respect to collections of web content and, in particular, blog content, and what policies have already been developed, on an institutional, national, and international level to address the issues?
2. What rights management capabilities might the BlogForever repository be able to provide (e.g. with respect to identifying rights management opportunities at key points in the digital information life cycle, authentication and authorisation technologies, and metadata assignment capabilities)?

6 http://creativecommons.org/
7 http://www.gnu.org/licenses/gpl.html
8 ccREL - http://wiki.creativecommons.org/CC_REL
9 ODRL - http://odrl.net/
10 http://scorm.com/
11 Advanced Distributed Learning - http://www.adlnet.gov/
12 http://registeredcommons.org
13 http://www.safecreative.org
14 http://www.sherpa.ac.uk/romeo/
15 http://corecopyright.org/2009/12/03/copyright_ip/

3. Is there a common conversation among experts about rights management that might provide insight to groups involved in blog content management and supporting technologies?
4. What approaches for rights management might be developed in the future?

We aim to aid future BlogForever repositories in answering these questions.

1.2 Contribution of this report

This report describes guidelines developed within the BlogForever project for designing rights management policies and procedures that support the preservation of weblogs. The deliverable is intended to

- inform you of the aspects of digital rights that need to be considered and approaches that might be adopted to build a policy suitable for you and your organisation, and,
- suggest practical ways to enable you to use the BlogForever platform while being cognizant of copyright laws, the intellectual property rights of content owners, or other rights-associated laws.

We have identified the main barriers associated with rights management as: intellectual property Rights (e.g. copyright, licensing, trademark, trade secret, patents), privacy (e.g. data collecting, data sharing and data protection), and legality (e.g. possession of illegal content, liability for defamation).

Our recommended approach to curators, librarians and archivists is not that they undertake extensive study of their local laws, or seek costly legal advice as a first resort, but use this deliverable to gain enough familiarity with rights management risks and issues, insofar as they affect your ability to use BlogForever, and insofar as they relate to your intended target blog collections. From that point, the suggested actions and treatments in this deliverable are a matter of mitigating those risks.

This document has two primary functions in assisting those who will manage future BlogForever repositories. Firstly, we intend to sensitise the reader to which issues of rights management are most critical for institutions that plan to engage in the archival of weblogs (addressing questions 1 and 2 of Section 1.1). Second, we refer to several strategies that the BlogForever project has identified as helpful for developing rights management policies that are appropriate for the location, resources and needs of those who will utilise BlogForever for developing repositories and express what is possible, within the current software platform, to address rights management (addressing question 3 of Section 1.1).

1.3 Disclaimer

Our aim with this document is to support the reader in acquiring a better understanding of some of the issues related to rights management and to provide them with quality information that can be used for making informed decisions.

It is not within the scope of this document to provide a comprehensive list of all relevant Copyright legislation, and even so it would still not address copyright problems in the digital realm.

It should be noted that the authors of this deliverable are not lawyers. They are not qualified to give legal advice. The report is intended as a guideline only.

1.4 Structure of the document

This report is structured to respond to the questions of Section 1.1. More specifically:

1. We respond to question 1 of Section 1 by expressing explicit DRM issues, offering examples of what challenges arise in relation to these issues, within the web, and, especially, the blog context and how other existing archives have approached the challenge (Section 2),
2. We respond to question 2 of Section 1.1 by highlighting rights management actions that might be implemented at different points of the digital information life cycle, in particular, in relation to the functional entities of the OAIS model (Section 3)
3. We respond to question 2 of Section 1.1 by describing functionalities of the BlogForever spider and platform that are designed to support digital rights management (Section 4), and, by providing concrete examples of how right metadata might be catalogued (Section 5).
4. We respond to question 3 of Section 1 by presenting interviews with experts in the field as a glimpse into the on-going conversation about the dilemma existing between digital preservation and rights management (Section 6).

Each section provides recommendations based on the findings. This will be brought together in Section 7 to offer our conclusions with regard to the objectives and aims of Task 3.3. Question 4 of Section 1 will also be addressed in this final section.

2 Issues Related to Rights Management

The range of areas that one could cover in relation to legislation, precedents and practices in digital rights management is extensive and the depth to which one could go within each topic is overwhelming. Here, the attempt is not to present an exhaustive study but to present a broad picture of the landscape to help future curators of digital materials navigate through different issues related to rights management.

The investigation of the partners in BlogForever shows that there are four main domains of concern: intellectual property (e.g. copyright, trademark, design, and patent), privacy (e.g. personal information, statutes, data protection), content associated with legal conflict (e.g. illegal content and activity, defamation), and agreements and licenses (agreements between content managers and content users that impose contracts on how and by whom content can be used). In the following sections, we will explore the issues related to the four areas that might need to be considered by curators of digital materials coming from the web and, more specifically, from blogs.

2.1 Intellectual Property Rights

Intellectual property relates to claims on an expression of an idea. It is not so much an ownership of the idea itself but ownerships related to expressions, representation, and implementations of an idea. There are numerous types of intellectual property but perhaps the best known are those that can be mapped to notions of copyright, trademark, design, and patent. Copyright governs the access, copy, and distribution rights associated to expressed pieces of work, trademarks are signs that "distinguish your goods and services from those of your competitors"[16] (this could also be called a "brand"), designs refer to "the way an object looks: its shape, its visual appeal"[17], and patents "protect the features and processes that make things work"[18].

Intellectual property is closely related to concepts of licenses and trade secrets. Licenses are not so much a description of the intellectual property ownership but expressions of "a partnership between an intellectual property rights owner (licensor) and another who is authorized to use such rights"[19]. It prescribes what the latter can do with the intellectual output in question. Trade secrets are governed by confidentiality laws based on non-disclosure agreements before information is shared.

There could, in fact, be several layers of licenses associated with the same content and/or idea, for example, by having a license agreement between the copyright owner and the repository and also between the end-user of the content and the repository in charge of distribution. These are likely to influence one another.

In the context of information on the web and blogs, all of these may come into play: for example the design, implementation, logos and content of blogs could be captured which might be protected by trademark, design, patent, confidentiality agreements, licenses and/or copyright. It is essential that a repository avoid conflict through openness and by acquiring the permissions to include these in the collection wherever possible.

With trademarks, designs and patents, the restriction is mostly on the way it is used (e.g. using trademark or design for sales, endorsement, or misrepresentation). Confidentiality laws that govern trade secrets come into effect through agreement. Licenses and copyright, however, are a fundamental barrier to all preservation actions, as crawling information on the web in itself is, strictly speaking, a breach of copyright or license unless prior permission had been obtained from

16 http://ww.ipo.gov.uk/types/tm/t-about/t-whatis.htm
17 http://www.ipo.gov.uk/types/design.htm
18 http://www.ipo.gov.uk/types/patent.htm
19 http://www.wipo.int/sme/en/ip_business/licensing/licensing.htm

the copyright owner or license holder. In the following we try to capture the current views on copyright and licensing with respect to digital content, and illustrate what the challenges are especially with respect to information on the web and blogs.

2.2 Copyright

To quote from the Joint Information Systems Committee[20] report on IPR: "For public bodies, understanding Intellectual Property Rights (IPR) and licensing is essential to their role as a provider, aggregator and/or publisher of publicly funded digital content"[21]. As a consequence the Joint Information Systems Committee (JISC) and the Strategic Contents Alliance (SCA) has developed an IPR toolkit[22] for understanding the basics of IPR, licensing, orphan works, digital economy act, and the use of third party content[23]. This resource deals mainly with a special case of IPR known as copyright which pertains to who can:

1. Make copies of the work;
2. Create new works based on the original (derivative works); and,
3. Distribute the work by sale, transfer of ownership, rental, lease, or lending"

Efforts towards international harmonisation of copyright laws has been going on as far back as the Berne convention in 1886[24]. There have been many treaties proposed towards the same end since implemented by the World Intellectual Property Organisation[25]. In the European Union alone there have been many implementations towards harmonisation[26], for example, related to:

- satellite broadcasting and cable retransmission (1993),
- fees payable for trademark and design (1995),
- database rights (1996),
- conditional access service, biotechnological inventions, and harmonisation of design laws (1998),
- community patents (2000),
- copyright in knowledge economy, resale rights, information society (2001),
- community design, and piracy (2002),
- internal market and enforcement of IPR (2004),
- musical works (2005),
- term of protection, and rental and lending rights (2006),
- patents (2007),
- exceptions and trademark (2008),
- computer programs, trademark law harmonisation, and enforcement (2009), and,
- orphan works (2011).

However, as much as harmonisation has been attempted, the final decision is still grey and the legislation has yet to catch up fully with the way we use information now on the web and, in particular, on blogs. For example, the legislative framework of the copyright of content does not have definite laws or regulations for "permitted" usage. The Copyright Act sets out four factors for courts to look at (17 U.S.C. § 107[27]):

20 http://www.jisc.ac.uk
21 http://sca.jiscinvolve.org/wp/allpublications/ipr-publications/
22 http://www.jisc.ac.uk/publications/programmerelated/2009/scaiprtoolkit
23 http://www.web2rights.com/SCAIPRModule/rlo1.html
24 http://www.wipo.int/treaties/en/ip/berne
25 http://www.wipo.int/copyright/law/
26 http://europa.eu/legislation_summaries/internal_market/businesses/intellectual_property/index_en.htm
27 http://www4.law.cornell.edu/uscode/17/107.html

1. **The purpose and character of the use.** Transformative uses are favoured over mere copying. Non-commercial uses are also more likely to be permitted.
2. **The nature of the copyrighted work.** Is the original factual in nature or fiction? Published or unpublished? Creative and unpublished works get more protection under copyright, while using factual material is more often permitted use.
3. **The amount and substantiveness of the portion used.** Copying nearly all of a work, or copying its "heart" is less likely to be fair.
4. **The effect on the market or potential market.** This factor is often held to be the most important in the analysis, and it applies even if the original is given away for free. If copied work is used in a way that substitutes for the original in the market, it's unlikely to be a fair use; uses that serve a different audience or purpose are more likely fair. Linking to the original may also help to diminish the substitution effect. Note that criticism or parody that has the side effect of reducing a market may be permitted because of its transformative character. In other words, if the criticism of a product is so powerful that people stop buying the product, that doesn't count as having an "effect on the market for the work" under copyright law.

2.2.1 Copyright and the digital dilemma

In our view, copyright law has not yet caught up with the realities of the way that digital content is created, shared, transmitted and curated. To put it simply, all the things you will need to do with BlogForever – crawling digital content, storing it, migrating it, preserving it and rendering it – all involve making copies.

Copyright law does its best to prohibit the copying of original material. web archiving, and the BlogForever platform, embrace the act of copying, firstly when crawling the original blog data, secondly when developing a digital surrogate from the original data, and thirdly in making this surrogate available on the Internet, which thousands of users can then access and copy onto their own computers. Further copying actions are inevitably involved in the digital preservation process.

If you cannot make copies of a blog, then you cannot crawl it, copy it into your repository, make backups, make dissemination copies, nor perform transformation or migrations for preservation purposes.

"The glory of digital items is that they can theoretically be accessed from anywhere, and by multiple simultaneous users. But copyright law hasn't quite caught up to accommodate the digital environment and allow us to (legally) use and preserve digital items in the full capacity that the medium allows."[28] (Megan Amaral)

In the UK at least, the Hargreaves Review of 2011[29] may change this situation to some extent. In digital preservation, making copies is a fundamental part of what a digital archive does. For some time, digital librarians and archivists have been concerned that copyright laws are being violated by making archival copies of digital objects.

The current problem, as Hargreaves describes it, is that there is a copyright exception in force for archivists, but (a) it has serious omissions, e.g. audio and film material; (b) the wording is vague; and (c) the exception doesn't apply to all institutions. To make preservation easier, the Hargreaves reforms will attempt to widen the preservation exemption: allow it to cover more types of content (in fact anything that is copyrighted), and apply it to more memory institutions. At the same time the review will do everything possible to keep the interests of copyright holders protected. However, in the UK, the law on copyright hasn't changed yet.

28 http://easydigitalpreservation.wordpress.com/?s=copyright
29 http://www.ipo.gov.uk/types/hargreaves/hargreaves-copyright.htm

Our recommended treatments for Copyright issues are outlined in Section 3 of this deliverable. The general trend of our recommendations is that it is good practice to negotiate with your target bloggers, identify the rights holders in advance, and advise them of your intent to crawl and publish their archived blogs. In other words, always seek permission from the owners. This is a very good strategy for mitigating copyright and IPR risks.

Through the use of licensing, permission agreements, and Creative Commons, then successful results can often be achieved without recourse to expensive legal advice, and you will be able to crawl, preserve, and publish your archived blogs in BlogForever.

If this strategy fails, there are also restrictive treatments that can be applied, including notice and take-down policies, restricting access to a geographical location, and even applying digital protection measures to digital objects. However, in the spirit of sharing and openness, our recommendation is that this restrictive approach should only be considered as a last resort.

2.2.2 Two views at the extreme ends of digital rights

Here we try to illustrate two extreme contrasting positions that might arise as a result of laws in different regions and/or contexts.

1. Digital rights cannot be copied for any purpose – without written acceptance from any owners of the content or publishers.

In 2012, the Danish Supreme court stated that capturing and republishing more than 11 words from news sites was not allowed without an agreement[30]. In contrast to this, the UK Supreme court ruling April 2013 stated that Internet users do not require permission to browse and view copyrighted material on web pages[31]. However, the court still wants the Court of Justice of the EU (CJEU) to investigate and clarify these matters[32].

2. Anyone publishing on the internet is automatically accepting this to be indexed and republished according to fair use upon the internet. Such fair use is based upon global accepted standards for internet behaviour, and cannot be limited by local legislation – since content can be published from a server in one country and displayed in all countries globally.

Since Internet activities are in constant conflict with these scenarios. Legal conflicts are increasing and could be exploding if anyone decided to follow upon any potential legal conflict detected. For example, Google is constantly in legal battles in several countries (e.g. with France[33] and Germany[34]), especially in Europe. However, there are high numbers of vendors using Google as a model, and that could be found easily and pursued for same legal conflicts.

Different legal jurisdictions define IPR of content and publishing on the net differently. For example, the term “fair use” was coined early on in the US and is not as harmonised in European courts. There is an ongoing discussion about the difference between the US and the EU, but the differences are considered to be more of process and terminology rather than concept[35]. Since the laws have been made before the Internet emerged, the courts have been projecting old laws onto the

30 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:62010CO0302:EN:HTML
31 http://www.out-law.com/en/articles/2013/april/unauthorised-browsing-of-copyrighted-material-online-is-legitimate-says-uk-supreme-court/
32 http://www.worldipreview.com/news/meltwater-ruling-a-win-for-internet-users-but-case-heads-to-cjeu
33 http://www.reuters.com/article/2013/02/01/us-france-google-idUSBRE91011Z20130201
34 http://www.cjr.org/cloud_control/german_copyright_law_passes_lo.php?page=all
35 http://www.harbottle.com/copyright-exceptions-and-fair-use/

new technology and reality. For example, EU law defined copying database to be illegal and have tried to use this as definition on what could not be copied from the Internet[36]. However, it's not often possible to see if a web site is a database.

Copyright is becoming so complicated that one could easily be subject to a claim and a very costly legal process. In fact, a US court actually stopped publishers who are said to have raised hundreds of such copyright claims, seemingly to enable profitable settlements with most claims. In this court ruling however, it seems it wasn't clear if it actually had the rights they claimed[37].

2.2.3 Challenges for web archiving

The scenarios in this section are intended to illustrate the ambiguities that arise when applying the copyright law to the Internet context, and, especially, to the blog context. Blogs and social media network information are especially difficult in that the dynamic nature of content creation usually means that there are multiple authors of content, design, and identity associated to the same blog from several jurisdictions, and that these are constantly changing as a the blog evolves.

The scenarios are intended to draw awareness to the importance of communicating with, if not obtaining permission from, blog owners in relation to the fact that their content is being harvested before repository managers take liberties to capture it. To say the least, capturing, preserving, allowing access to blog content should be undertaken with some care. We describe seven topics of copyright and associated scenarios that illustrate the delicacy of rights management with respect to web and blog content.

1. "**Fair use"** to define what is acceptable use of 3rd party content. According to Wikipedia, "fair use" in US includes the right for search engines to copy and use the content of 3rd parties[38]. The concept of fair use does not exist in every country, and when it does exist it is not necessarily interpreted in the same way (e.g. see discussions of fair use as used in the UK[39] - this is sometimes termed "fair dealing"[40]). The global nature of use of information on the Internet introduces difficulties.

2. **Private use only.** Indexing and copying is only allowed for free services and private usage such as Google. However, Google is a commercial model and is used extensively by employees. There is no technical capability currently in place to limit use at work to private usage. However there are some movements in the UK to make private copying legal[41].

News Copyright holder organizations in Norway (Klareringstjenesten) and UK (NLA), are introducing policies on Google News as being acceptable without a license since it's a free and private-only service. In the court battle of Moreover vs Associated Press, however, Google was said to be driving more traffic than Moreover[42].

3. **Opt-in vs Opt-out for search engines**. Indexing with displaying links and short text snippets as part of search engines is considered acceptable given the availability of opt-out and robot.txt, as this considered part of the nature of Internet. In fact, in 2006, there is

36 http://ec.europa.eu/internal_market/copyright/prot-databases/index_en.htm
37 http://www.pcworld.com/article/244344/publisher_drops_copyright_claim_favors_fair_use.html
38 http://en.wikipedia.org/wiki/Fair_use
39 http://www.copyrightservice.co.uk/copyright/p09_fair_use
40 http://en.wikipedia.org/wiki/Fair_dealing
41 http://www.theverge.com/2012/12/21/3791352/uk-government-details-copyright-and-fair-use-revisions
42 http://paidcontent.org/2007/10/10/419-ap-sues-moreover-and-verisign-for-stories-copyright-infringement/

precedence of a court ruling in favour of Google in displaying cached pages[43]. In June 2013 German copyright issues has made Google change from opt-out to opt-in for German news sites[44], that is, sites are only indexed if owners have indicated that they explicitly allow Google to do so. However, if opt-in is to become required for search engines, there will be fundamental issues for coverage and may undermine the benefits of the Internet.

4. **Multiple ownership.** Who owns the copyright? If the blog author is the content owner, how can their ownership be interpreted in relation to the DRM policy of the blog platform (e.g. platforms such as wordpress.com, blogger.com)? For example, if the platform is using a ping server, they are actively sharing updates and new links with all spiders that subscribe to the ping server. This can be considered an active acceptance to be indexed and shared. But if the blog author has stated terms in a robot.txt restricting spiders from indexing the content, inconsistencies will arise which is an issue both for spiders and for the publishers.

The multiple layers of copyrights and multiple ownership can create a frustrating situation for end users. For example, a songwriter described how he ended up with copyright claims with YouTube[45] about his own songs[46]. In a supreme court ruling in Scandinavia, a buyer of a picture, bought from an agency holding the ownership rights obtained from the photographer, was taken to court for infringing rights because it was used without the consent of the person depicted in the photograph[47].

5. **Multiple legislation.** Which sets of laws should be applied when IPR owners are operating in different countries and the spider or platform is situated or hosted from yet another country. The end user may be situated in yet another country and handling legal conflict between all parties concerned can be difficult. One IPR owner might raise the case according to the local legislation, but it not clear how this will be taken by a legal entity outside the country, especially if it is not in line with the other relevant legislations.

Historically, it does not seem like claims have been raised frequently across regional boarders. There are also services available online offering to keep access limited to a single jurisdiction (e.g. BBC iPlayer[48] does not allow access to their content from outside of the UK[49]). Nevertheless, conflict with publishers from Thailand selling content in the US has been cited[50].

6. **Opt in and ping servers.** As a means of promoting blog content, blog authors and platforms publish their posts through ping servers. The ping servers hold all new URLs sent to them and feed this to any subscriber to the blog feeds. This is integral to the blog syndication activity. Nevertheless, there might be copyright terms prohibiting the harvest of this content. Unless such terms are expressed in a machine readable format or restriction is specified in robot.txt it is impossible to distinguish between content inviting harvest and that which is restricted. This undermines the broadcasting power of the blogosphere.

2.3 Licensing

43 Field v. Google Inc., 412 F. Supp. 2d 1106 (D. Nev. 2006) - See http://fairuse.stanford.edu/overview/fair-use/cases/
44 http://www.pcmag.com/article2/0,2817,2420854,00.asp
45 http://www.youtube.com
46 http://chriszabriskie.com/2013/04/how-i-end-up-with-youtube-copyright-claims-on-my-own-songs/
47 http://www.internationallawoffice.com/newsletters/detail.aspx?g=ce757358-3f3d-4bca-8747-906d82dd58d0
48 http://www.bbc.co.uk/iplayer/
49 http://iplayerhelp.external.bbc.co.uk/help/outside_the_uk/outsideuk
50 http://www.scotusblog.com/2013/03/opinion-analysis-justices-reject-publishers-claims-in-gray-market-copyright-case/

Based on the copyright and IPR regulations presented in Section 2.2[51], the archivist should have a clear understanding of the potential risks related to the copyright of the content that he is archiving. Therefore a careful risk assessment is advised as early as possible in the developmental process. In addition, it is important that the archivist or the repository manager to ensure that processes are in place to ensure that risk management is an ongoing activity, and that responsibility for undertaking this assessment, as well as developing and administrating methods of handling any risks identified, is clearly located within the staffing structure of the repository.

Creative Commons[52] licenses provide several copy and share licenses complete with legal code, computer code, and a human-readable declaration as well as a visual representation to let others know that they're invited to copy and share. Therefore curators, users, providers of the content can all easily determine whether attribution is required, and whether commercial use, or modifications are allowed. If someone wants to do more than is permitted by fair use or the terms of your license, they can still contact the blogger/author for permission.

As far as existing practice (Charlesworth 2009) in addressing copyright issues are concerned, there is a degree of support among stakeholders in all types of digital repositories for the adoption of clear and concise copyright licensing options like those provided by the Creative Commons (CC) project. What is also clear, however, is that:

- using CC licenses still requires at least a basic understanding, on behalf of both archivist and authors, of how copyright licensing works, and what is being granted (or not) by the author, and such knowledge is by no means universal;
- in many cases the IP rights in archives may be vested in third parties other than the repository manager; for the repository to make use of those resources may thus require the depositor to seek additional permissions;
- the license options available under the CC do not necessarily provide a complete solution to a repository's needs, e.g. if some depositors want more specific/restrictive terms;
- even if CC licenses (or variants thereof) are used, there remains the issue of how to deal with the results of the unintended or unsuspected incorporation of unlicensed third party material within archives.

As such, CC licenses does not always provide a solution for all deposit and access-related copyright issues arising in repositories (Korn & Oppenheim 2006). Depending on local or sectoral factors, repositories seeking to access archives may be better served by variants based on CC licenses or, indeed, entirely different licensing models.

It is important that there is a thorough assessment of those factors that will play a key role in aiding repository managers in choosing an appropriate licensing mechanism, from the beginning. Obtaining a viable set of quality digital objects through the repository's deposit process is a vital objective. It is important, therefore, that processes designed to facilitate copyright compliance, and to ameliorate risk, do not have the undesired consequence of deterring potential depositors.

It is essential for the repository managers to assess the factors that are likely to affect willingness to archive, and to tailor their processes accordingly, for example:

- a requirement on curators to create rights metadata for deposited materials would until recently have been seen as a negative factor in encouraging archiving material; however, increasing use of Web 2.0 technologies, such as 'tag clouds', may mean that archives creators/depositors are more willing to accept the benefits of metadata usage, and thus willing to accept some additional overhead to improve deposit processes;

51 Further information can be found in EFF site https://www.eff.org/issues/bloggers/legal/liability/IP

52 Additional information can be found in Creative Common's licensing page http://creativecommons.org/licenses/

- providing a small set of license choices from which depositors can choose will reduce confusion, but may also restrict the number of depositors who are able or willing to contribute under the sets of license terms available to them.

Part of this process will involve identifying areas in which a repository can enhance understanding through provision of a tailored range of information on licensing, and outreach mechanisms such as guidance and guidelines on IPR for depositors.

In addition to using explicit assignment of licenses, the Digital Millennium Copyright Act (DMCA - 17 USC § 512)[53], is another way for a repository to protect themselves from copyright liability, as service providers who "respond expeditiously" to claims that they are hosting or linking to infringing material are less likely to be sued. The DMCA does not make service providers liable if they do not remove content, but gives them a strong incentive to take the content down. Service providers who fail to remove content may lead to a strong incentive to make claims of copyright infringement. If a content publisher received a DMCA take-down notice, but he/she believes the material he/she posted does not infringe copyright, they have the option to counter-notify. Materials can be put back up after a counter-notification and still keep its immunity from liability. If harm results from an erroneous take-down demand, action can be taken to sue back (DMCA, section (f), 17 USC § 512[54]).

Additional relevant resource on the subject of protecting yourself against copyright claims is provided by the Digital Media Law Project (DMLP), hosted by Harvard University's Berkman Center for Internet & Society[55]. The IEEE guidelines for intellectual property rights for authors, readers, researchers and volunteers also provides a view on the publisher's point of view on content rights standards on the Internet[56].

2.4 Privacy

Apart from intellectual property rights, another big concern arising within the context of digital rights management that should not be taken lightly is the respect for privacy that a digital repository might be expected to uphold. We will use the term privacy, in the current context, to encompass concerns about surveillance, data protection, and statutory rights. These all involve establishing a clear approach to what *personal information* will be collected, shared, protected, and used. The issue of privacy can refer to the following:

1. "an individual's right to control the collection, use and disclosure of information about him or herself" (Henderson & Snyder 1999; DELOS 2007)
2. A "Privacy and Confidentiality Policy", or "a *policy* outlining the terms by which the organisation that manages the DL [Digital Library] will handle personal information on its *Actors"* (DELOS 2007; DL.org 2010)
3. A privacy policy describing how an institution "uses, collects, and shares information through the services"

The European Union Directive on Data Protection was published in 1995[57]. The newest amendments directive for data protection in the European Union was published in 2012[58]. International laws (e.g. those practiced in Australia[59], United States[60], Canada[61], New Zealand[62])are

53 http://www.law.cornell.edu/uscode/text/17/512
54 http://www.law.cornell.edu/uscode/text/17/512
55 http://www.dmlp.org/legal-guide/protecting-yourself-against-copyright-claims-based-user-content
56 http://www.ieee.org/publications_standards/publications/rights/index.html
57 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:NOT
58 http://ec.europa.eu/justice/data-protection/
59 http://www.oaic.gov.au/privacy/privacy-act/the-privacy-act
60 http://www.justic.gov/opcl/

also quite specific but tend to be regional. One of the strictest data protection/privacy laws are that of Republic of Korea, only recently introduced (Greenleaf & Park 2012[63]). The Korean model is relevant because the new law claims to conform to the European laws while being stricter. To contextualise the landscape, we summarise the key points of the new Act which is based on seventeen principles:

1. In the case of a dispute the responsibility of proof is on the processor not the one claiming the breach;
2. Only minimal amount of personal data collection necessary for the collection is allowed. Efforts to process data so that identity of persons can remain anonymous is required.
3. There should be no denial of service based on a person's refusal to provide information that is not legally required.
4. Processing of sensitive data can only take place after consent.
5. Alternative identification method other than the Residence Registration Number must be provided.
6. Strict limits on visual surveillance devices.
7. A Privacy Policy must notify users how to opt out of personal data collection.
8. Data subjects must be notified when data is collected from third parties.
9. Data can be disclosed to third parties only after consent.
10. If a sub-processing agency is employed data subjects must be notified.
11. Once the purpose of processing has been achieved, data must be deleted.
12. Suspension of data processing can be required by data subject.
13. Policy must be issued.
14. Privacy Officer must be appointed.
15. If data breach takes place, data subjects must be notified.
16. Detail security measures must be prescribed.
17. Data exports must be preceded by consent from data subjects.

The Korean authorities (Data Protection Commission; Korean Internet and Security Agency; Personal Information Dispute Mediation Committee; Ministry Public Administration and Security; and the Korea Communication Commission) have been shown to be conscientious enforcers of previous laws so the next years will serve to test whether such strict laws are practical solutions within the Internet environment.

2.4.1 What is "Personal Information"?

Privacy policies make reference to the collection of "personal information". "Personal information" means recorded information about an identifiable individual, including:

- information relating to the race, national or ethnic origin, colour, religion, age, sex, sexual orientation or marital or family status of the individual,
- information relating to the education or the medical, psychiatric, psychological, criminal or employment history of the individual or information relating to financial transactions in which the individual has been involved,
- any identifying number, symbol or other particular assigned to the individual,
- the address, telephone number, fingerprints or blood type of the individual,
- the personal opinions or views of the individual except if they relate to another individual,
- correspondence sent to an institution by the individual that is implicitly or explicitly of a private or confidential nature, and replies to that correspondence that would reveal the contents of the original correspondence,

61 http://www.www.priv.gc.ca/leg_c/leg_c_a_e.asp
62 http://www.legislation.govt.nz/act/public/1993/0028/latest/DLM296639.html
63 http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2120983

- the views or opinions of another individual about the individual, and,
- the individual's name where it appears with other personal information relating to the individual or where the disclosure of the name would reveal other personal information about the individual.

This list of examples of personal information is not exhaustive. To qualify as personal information, the information must be about the individual in a personal capacity. As a general rule, information associated with an individual in a professional, official or business capacity may not be considered to be "about" the individual. Even if information relates to an individual in a professional, official or business capacity, it may still qualify as personal information if the information reveals something of a personal nature about the individual. Finally, to qualify as personal information, it must be reasonable to expect that an individual may be identified if the information is disclosed.

2.4.2 Privacy in Web Archiving

All institutions engaging in the archiving of digital materials must consider the issue of privacy in terms of personal information obtained about users of content (or producers of content), as described in the previous section. Additionally, archiving institutions must have policies on handling usage data and user content posted to the services. User data refers to the IP address of a user, the URL request, browser type and the date and time of a request. User content posted to the services refers to any comments submitted, saved settings, or saved searches that can be connected to a particular user (by use of registration pages or other means). In Tables 2.1 to 2.5, we provide some examples of how the privacy issues mentioned above are dealt with by specific archives, and how this information is collected and used.

Type	**Action (When)**	**Description (How)**	**How information is used**
Usage Data (your IP address, URL request, browser type, and the date and time of a request)	DPLA collects this information when you access the Services, including when you set up an account, upload information, or browse, read, or download information from the Services.	may use cookies or other automated mechanisms	DPLA uses the information that it collects to deliver and improve the Services, administer mailing lists and online communities, and other activities related to the provision of the Services. DPLA may retain all data and content collected through the Services for restorative, archival, or research purposes. DPLA may share the information it collects through the Services with third party service providers as necessary to provide or improve the Services.
Personal Information (e.g. email address)	In order to access certain parts of the Services, DPLA may require you to provide personal information, such as your email address.	The user provides it by registering for an account or otherwise posting, sending, or uploading the information to the Services.	DPLA may also share certain information it collects with its research partners. In addition, although DPLA will not publicly share personal information unless you choose to make that information public, DPLA may publicly share aggregated data or statistics related to the Services that may

Type	Action (When)	Description (How)	How information is used
			include anonymised user information or usage statistics. Finally, DPLA may share information when responding to a request from law enforcement, or to prevent malicious use of the Services.
User Content that You Post to the Services	Certain parts of the Services allow you to upload or post comments or other content ("User Content"). Likewise, the Services also allow registered users to save searches ("Saved Searches") and to save and share lists of items found through the Services ("Saved Lists").	User Content may be publicly available, viewable to others, and may appear in search results on third-party search engines.	

Table 2.1: Digital Public Library of America[64]: approach to issues of privacy[65]

Type	Action	Description	How information is used
Usage Data	Because the Archive uses standard Web logging in its Web servers, our Web server may automatically recognize the domain name of each Visitor, each Visitor's IP address, what Web page the Visitor requests, and the time of the request, along with a variety of information supplied by the visitor's browser.	Web logs	The Archive may disclose any information it collects from Users if the Archive believes in good faith that such action is reasonably necessary to enforce its Terms of Use or other policies, to comply with the law, to comply with legal process, to operate its systems properly, or to protect the rights or property of itself, its Users, or others.
Personal Information	The Archive may collect the email addresses and messages of those who communicate with it via email or who enter email addresses in forms. The Archive may collect personally identifying information when a Researcher registers for access to the Collections, including the Researcher's name, address,	The Archive may use "cookies" to track Users' activities on the Site and in the Collections	The Archive may transfer the information on its machines, including personally identifying information, into the Collections. The Collections are made available to researchers and may be made available on the Site, or provided to third parties, for any use, without limitation.

64 http://dp.la/
65 http://dp.la/info/terms/privacy/

Type	Action	Description	How information is used
	telephone number, and email address, and the Researcher's proposal for using the Collections.		

Table 2.2: Internet Archive[66]: approach to issues of privacy[67]

Type	Action	Description	How information is used
Personal Information	The Archive collects personal information when you register with the Archive, when you use the Archive services, when you visit the Archive. The Archive may collect personally identifying information when a Researcher registers for access to the Collections, including the Researcher's name, address, telephone number, and email address, and the Researcher's proposal for using the Collections.		The Archive uses information for the following general purposes: - fulfil your requests services, - improve our services, - contact you, conduct research, and provide anonymous reporting for internal use.
Usage Data	The Archive collects receives and records information on our server logs from your browser, including your IP address, the Archive cookie information, and the page you request.	Server logs, cookies	Cookies are pieces of information that the Archive will transfer to your computer's hard drive through your browser to enable the Archive's systems to recognise your browser. Cookies also enable the Archive to gain information about the use of its Website and to enhance the Website accordingly to the preferences of the users.

Table 2.3: Internet Memory Foundation[68] and European Archive[69]: approach to issues of privacy[70]

Type	Action	Description	How information is used
Usage Data	When you visit this website, we collect website usage		

66 http://archive.org
67 http://archive.org/about/terms.php
68 http://internetmemory.org/en/
69 http://www.europarchive.org/
70 http://www.europarchive.org/terms.php

Type	Action	Description	How information is used
	information and information about your computer and internet connection, including your computer's IP address, the type and version of browser (such Internet Explorer 7 or Firefox 3.6) and operating system you use, your internet domain and, if you arrived at nationalarchives.gov.uk via a link from another website, the URL of the linking page.		

Table 2.4: The UK National Archive[71]: approach to issues of privacy[72]

Type	Action	Description	How information is used
Usage Data		The cookies used by British Library websites do not contain any of your personal information, and we cannot use them to find out who you are.	We use cookies to analyse how visitors use our website, for example, to identify which pages on our site are the most popular or to allow you to store your preferences. we use cookies to remember that you are logged in; record items placed in shopping baskets; and to remember search terms and search results.
Usage Data	This website uses industry standard analytics packages which automatically gather information on visitors to our website. This information is stored anonymously in server log files.	server log files	It does not identify individual users and is used only for website administration and analysis of website usage and trends.

Table 2.5: The British Library[73]: approach to issues of privacy

It is important for the user to be able to trust a platform/service in order to use it or give his consent to crawl/store his content. In a privacy policy it should be clear what information the platform/service collects for its users, how it uses them and if there are cases where the information will be disclosed and under which circumstances these cases will occur.

For example, when a platform/service uses cookies then the user should be informed that if he sets his browser to refuse cookies then certain features may not function properly without the aid of cookies. Also, the user should be informed about what will happen in case the company goes out of business or enters bankruptcy and if there is the possibility to delete his account and what happens

71 http://www.nationalarchives.gov.uk/
72 http://www.nationalarchives.gov.uk/legal/privacy.htm
73 http://www.bl.uk/

when he deletes his account. Moreover, the user should be aware that some factors such as unauthorized access, hardware or software failure, may compromise the security of user information, what actions he can do to prevent unauthorized access to his account and personal information such as to log out of his account after finishing using the platform/service and to maintain the security of his password by choosing and protecting a strong password. Another issue that should be addressed is transfer of information. What happens in the case that there is collaboration with another platform/foundation? And, how will user information be disclosed in response to court orders, or legal process? Finally, the user should know if changes may happen to current privacy policy and how he will be informed about them.

2.5 Defamation and Illegal Content and Activity

In 2003, the UK parliament introduced a bill for a Defamation Act[74]. This bill was proposed to "ensure that a fair balance is struck between the right to freedom of expression and the protection of reputation". The principles of this bill are that it:

- "includes a requirement for claimants to show that they have suffered serious harm before suing",
- "removes the current presumption in favour of a jury trial",
- "introduces a defence of 'responsible publication on matters of public interest'",
- "provides increased protection to operators of websites that host user-generated content, providing they comply with the procedure to enable the complainant to resolve disputes directly with the author of the material concerned",
- "introduces new statutory defences of truth and honest opinion to replace the common law defences of justification and fair comment".

In the United States the laws for defamation tend to be much less plaintiff-friendly than in the European Union and/or the Commonwealth countries, due to the enforcement of the First Amendment[75]. It also differs across different states.

Illegal content or activity pertain to issues such as pornography (which may be banned in some countries), child sexual abuse material (including child pornography), online grooming, and hate speeches that incite prejudices against an identifiable group of people.

In places like Korea, they are striving towards policies that support freedom of speech. However, this trend conflicts with the National Security Act that has been in place since 1947 when the country split into two political groups. Since the Korean war in 1950 Korea has officially been at war. These situations affect the way people look at defamation and illegal activity and content.

The only way that conflict can be prevented, if at all, is through making explicit what material cannot be accepted into the archive, library and/or repository right at submission stage of the blog and to get confirmation from the content providers that they are adhering to the rules of the repository, and provide mechanisms for users to raise issues and resolve issues through the repository managers and curators of the digital materials.

2.6 Conclusions

There are a lot of grey areas when it comes to resolving issues of intellectual property, copyright, licenses, defamation and illegal content and activity. The best practices in defining digital rights management policies rely on common sense. For example, high profile organisations are more

74 http://services.parliament.uk/bills/2012-13/defamation.html
75 http://www.archives.gov/exhibits/charters/bill_of_rights_transcript.html

likely to be targeted for legal action than those that are low profile. Cases where commercial benefits play a substantial role are likely to attract more disputes (nobody likes other people making money off something they produced).

Here we provide a basic set of approaches to consider in developing digital rights policies:

- express clearly what materials are being collected as part of your collection and repository function and what will be used and how it will be used;
- notify blog owners what is being harvested, to indicate that efforts are being made to seek permission;
- state what materials are not allowed for submission and get agreement from the submitter that no such materials have been knowingly included;
- make it easy for users to flag up material that might be breaching copyright, licenses, privacy;
- make it easy for users to flag up materials that might be involved in defamation and/or illegal activity;
- respond in a timely manner to user requests and make it easy for involved parties to resolve any disputes: e.g.
    - ✓ communicate as soon as possible with the reporter to indicate that the matter is being investigated,
    - ✓ provide a way for the content provider to communicate with the claimant for the purpose of resolving matters of dispute,
    - ✓ quarantine items under investigation;

- make every effort to:
    - ✓ communicate repository activity,
    - ✓ implement acquisition of rights metadata at different points of repository processes and use accepted standards for expressing them,
    - ✓ retrieve agreements and licenses that are already in place whenever possible to accompany content,
    - ✓ make explicit that the above actions are taking place,

- keep abreast of conversations on digital rights to monitor changes in a legal landscape for web and blog archiving and keep a program for periodic risk management.

The discussions on licenses in Section 2.3 and defamation and illegal activity in Section 2.5 especially highlight digital rights management as risk management activity. This might be effectively investigated using tools such as the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA[76]).

In the rest of this report, we will provide approaches to supporting these activities. For example, we will revisit the repository digital information life cycle (Section 3.1), to identify opportunities for digital rights management, and present recommended practices within this workflow (Section 3.2). This discussion will be followed by a discussion of processes already in place within the BlogForever spider and repository that might help to support digital rights management, and that might help identify methods to automatically retrieve and make available license agreements and copyright terms already expressed in target blogs (Section 4). The report, further, makes recommendations related to cataloguing right metadata (Section 5), and taps into ongoing conversation in digital rights management through selected interviews with experts (Section 6).

76 http://www.repositoryaudit.eu/

3 The Digital Information Life Cycle and DRM Best Practices

This chapter suggests opportunities for dealing with rights management issues and risks when working with BlogForever. We begin with a suggested workflow based on the OAIS model (CCSDS 2002), with some suggested treatments related to that workflow. The second half of the chapter is a survey of best practices for rights management in web-archiving, based on existing practices from various international repositories. These are aligned with further opportunities which are not explicitly identified within the OAIS workflow, for example a Selection stage.

3.1 Using a workflow for rights management

This section describes possible rights management opportunities in the OAIS workflow, with respect to five of the six functional entities: Ingest, Storage, Access, Policy, Administration.

We recommend that users of the BlogForever service consider building a defined workflow for the repository as the best way to achieve rights management. Rights management is not defined within the system, for reasons already stated in the introduction. Our recommendation therefore is that the user should define their own customised rights management workflow.

One possibility for doing this is the OAIS model[77]. OAIS has been referenced extensively in deliverable BlogForever: D3.1 Preservation Strategy Report (September 2012) of this project. OAIS is principally a means to ensure long-term preservation of digital resources in a repository environment, and the deliverable proposed an OAIS-like repository workflow that is suitable for weblogs. However, a curator using BlogForever can also think about responses to their legal issues in relation to the OAIS workflow. We propose stages in the repository workflow where it would be possible to intervene, and the form these interventions would take. These can map directly to particular OAIS functions.

3.2 Rights management in OAIS

Below are some suggestions for rights management opportunities using the OAIS Model.

OAIS proposes six *functional entities*: Ingest, Archival Storage, Data Management, Administration, Preservation Planning, and Access. Within each functional entity, OAIS describes more detailed *functions* to manage the information flow. In our report, we will use these functions as potential opportunities in a repository for rights management interventions. To illustrate the point visually, we use copies of diagrams taken from the OAIS Model with added callout boxes. In the text below, the functions are expressed in italics.

Our report refers to the concept of Information Packages as described in OAIS; an Information Package in BlogForever is a crawled blog (the data), plus metadata about the Blog. There are three types of Information Package: Submission, Archival, and Dissemination, which we refer to as SIP, AIP and DIP. For further details on how SIP, AIP and DIP relate to BlogForever see the BlogForever Deliverable D3.1, Preservation Strategy Report (September 2012), pp 131-133.

We also use the OAIS term "Producer" to designate the *creators of content* (bloggers); and the OAIS term "Consumer" for *users of the archived blogs* (such as academics, journalists and researchers).

77 ISO 14721:2003 http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

OAIS has the following to say about Copyright and Rights Management, which it correctly regards as part of the responsibility 3.2.2 "Obtains Sufficient Control For Preservation":

"Copyright implications, intellectual property and other legal restrictions on use: An Archive will honor all applicable legal restrictions. These issues occur when the OAIS acts as a custodian. An OAIS should understand the intellectual property rights concepts, such as copyrights and any other applicable laws prior to accepting copyrighted materials into the OAIS. It can establish guidelines for ingestion of information and rules for dissemination and duplication of the information when necessary. *It is beyond the scope of this document to provide details of national and international copyright laws.*" (Our emphasis) [78]

Beyond the above statement, there are very few explicit suggestions in OAIS for how to perform rights management activities or how to achieve legal compliance. Therefore this section of D3.3 represents our *interpretation* of how rights management would be possible in the OAIS framework.

Of the six principal OAIS functional entities, the only one not used in this report is Data Management, which in OAIS describes using a repository database to co-ordinate and record actions. It has been excluded because:

1. There are no rights management opportunities in this functional entity, other than the possibility to update rights metadata
2. The Data Management function is already being performed by Invenio's databases

3.2.1 Opportunities in the Ingest functional entity

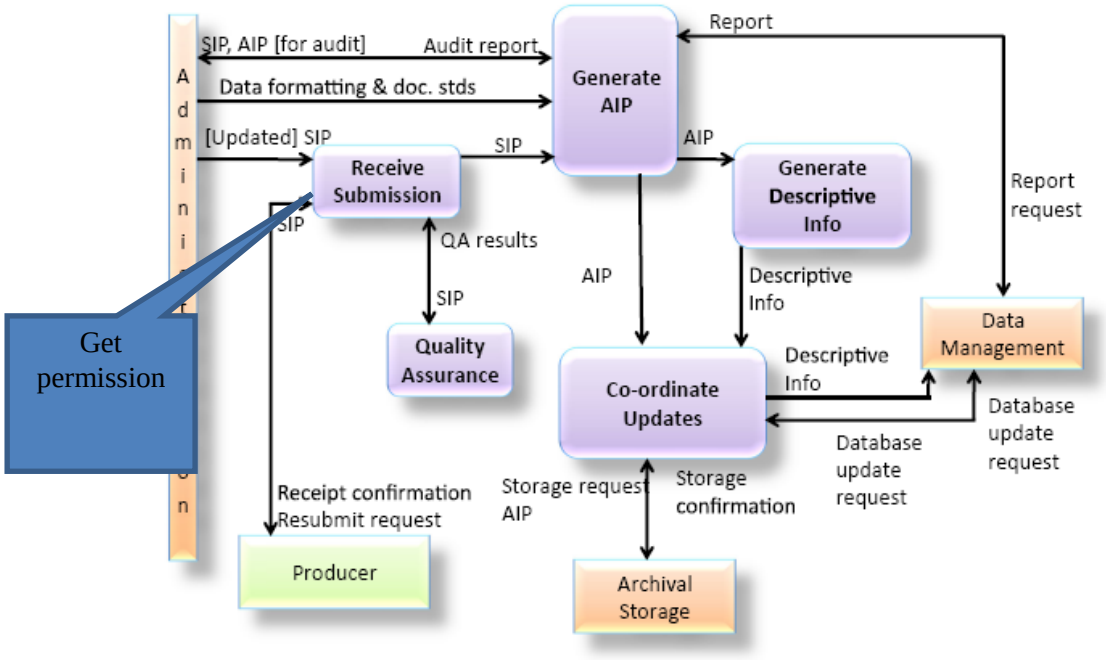


Figure 3.1: Functions of the Ingest Functional Entity (Source: OAIS June 2012, page 4-5).

Function: Receive Submission

78 Open Archival Information System CCSDS 650.0-M-2, June 2012 (The Magenta Book), page 3-2.

Action: Get permission to harvest / copy / store / republish the blog

Description: Ask bloggers / submitters to sign up to a deposit agreement. You can ask what types of access or usage are permitted by the rights owners.
There are numerous opportunities to negotiate with blog owners and seek permission to crawl; identify copyright owners; create and draft licenses, or consent forms; and other actions to mitigate any risks associated with rights. In many cases it is important to seek permission before you start crawling a blog.

Outcome / result: documented / recorded evidence of agreements that will be essential in case of any future disputes.

For examples of a Permission agreement, see section 3.4.1 below; for a suggested consent form, see sections 3.4.2 and 3.4.3.

Other outcomes are detailed in section 3.5.3, including the process of negotiating with rights owners, access restrictions, seeking copyright license, gaining permission from third-party rights holders, opt-out forms, and publishing statements about copyright implications.

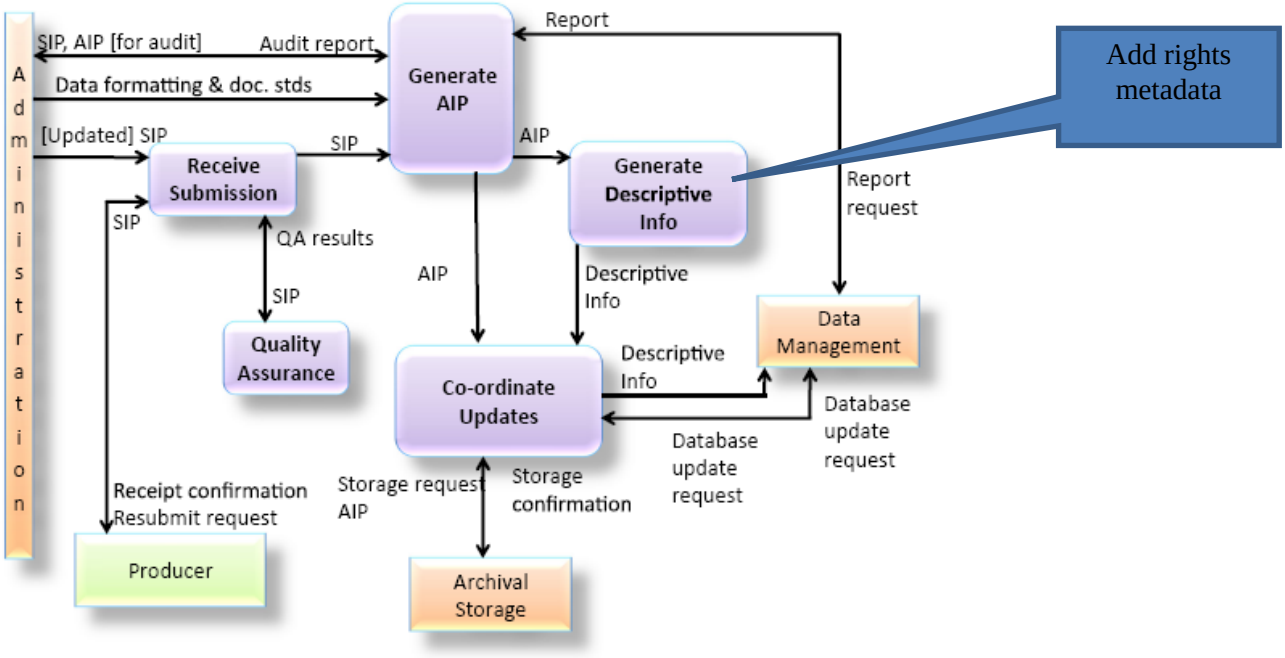


Figure 3.2: Functions of the Ingest Functional Entity (Source: OAIS June 2012, page 4-5).

Function: Generate Descriptive Information

Action: Add rights metadata to the SIP or AIP.

Description: This could be a statement from a blogger asserting copyright of their work; it might already have been captured in some form by the spider and so be part of the SIP. Depending on the schema you use, this information could be parsed in some detail. See the section on cataloguing rights metadata for further information.

Outcome / result: permanent records of rights information associated with the blogs you are harvesting in your organisation. Such records can be referred to in case of any disputes or issues. It is also possible to store the metadata in a database, thus enabling better management.

See section 5 on cataloguing rights metadata for further information, including examples and suggested cataloguing rules.

3.2.2 Opportunities in the Archival Storage functional entity

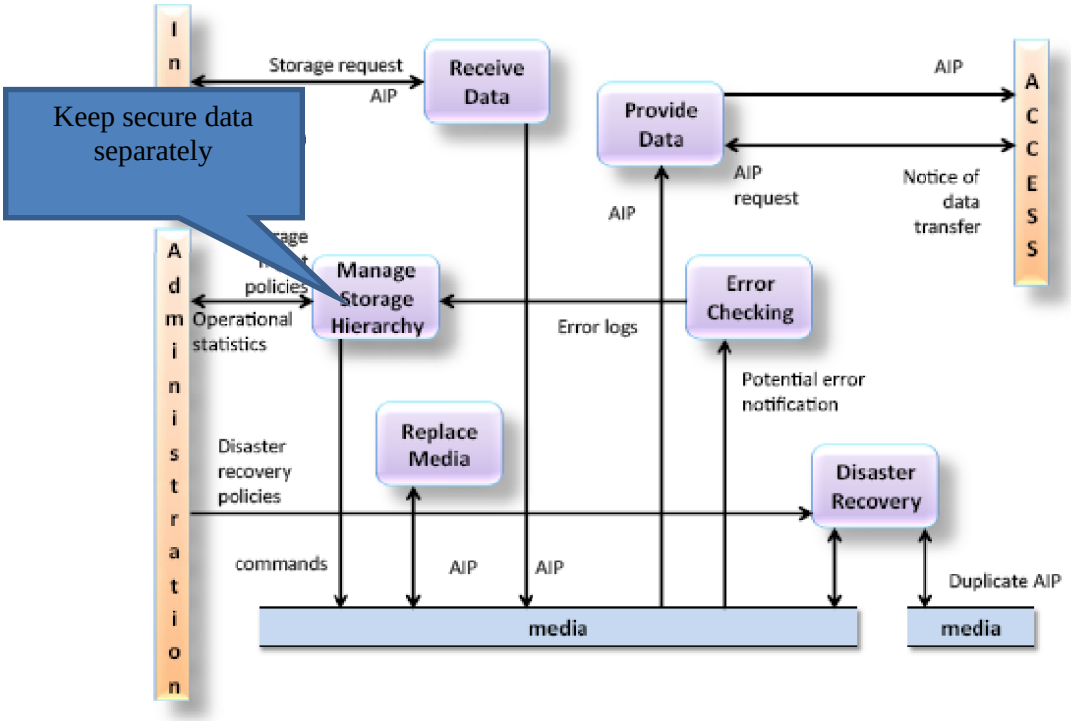


Figure 3.3: Functions of the Archival Storage Functional Entity (Source: OAIS June 2012, page 4-8).

Function: Manage Storage Hierarchy

Action: Store restricted AIPS separately.

Description: here, you can interpret the function to help manage legal uses. You can build a dedicated server (or partition of a server) for storage closed content, while another dedicated server serves accessible content. This highly secure part of the preservation system could only be accessed by archivists and curators, or administrators. Such an approach will be informed by your written policies on security.

Outcome / result: Sensitive blog data (if any) is stored in an appropriate secure fashion, thus minimising or eliminating any risks of broaching privacy laws or data protection acts. See section 2.4 for an understanding of privacy laws.

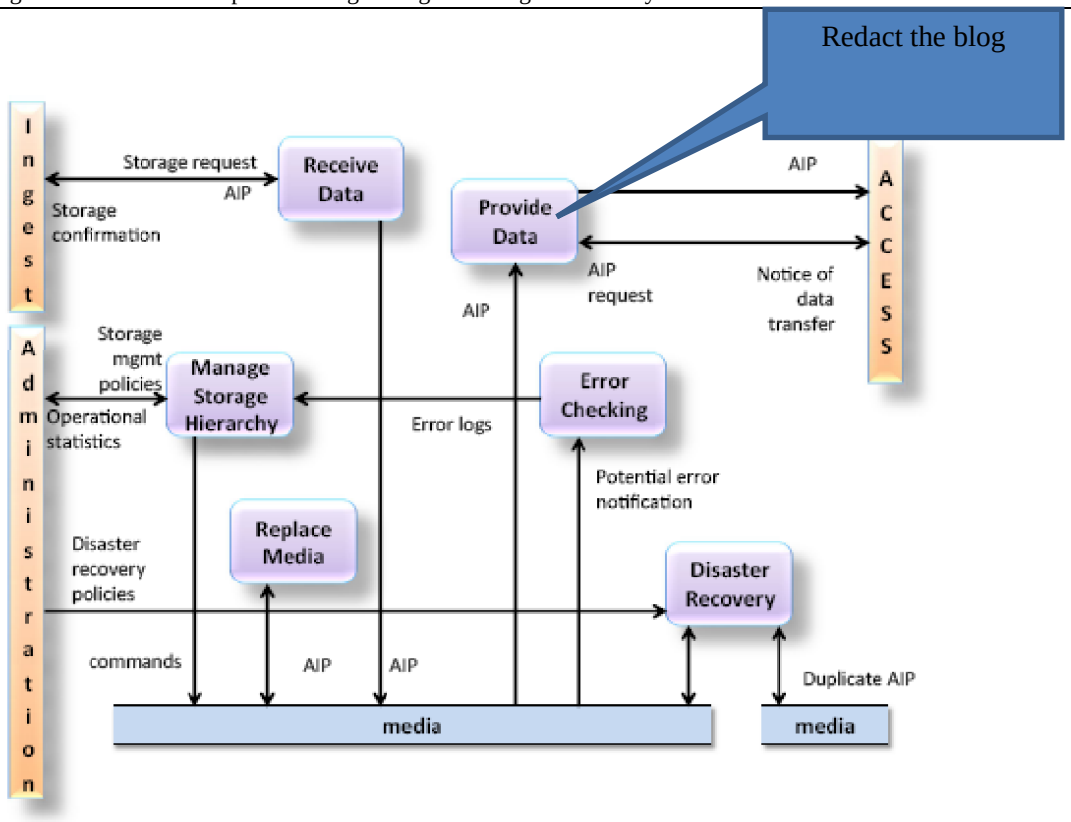


Figure 3.4: Functions of the Archival Storage Functional Entity (Source: OAIS June 2012, page 4-8).

Function: Provide data

Action: Redact AIP before Access

Description: This step in OAIS describes copying the AIP out of storage in response to an Access request. The Information Package isn't yet a DIP, but at this stage you could redact it and start to turn it into a redacted DIP. This could mean hiding words in text files, closing certain sections of the blog.

Outcome / result: This action ensures that sensitive blog data (if any) is never served or released to users, and thus protected from wrongful use at point of access.

3.2.3 Opportunities in the Administration functional entity

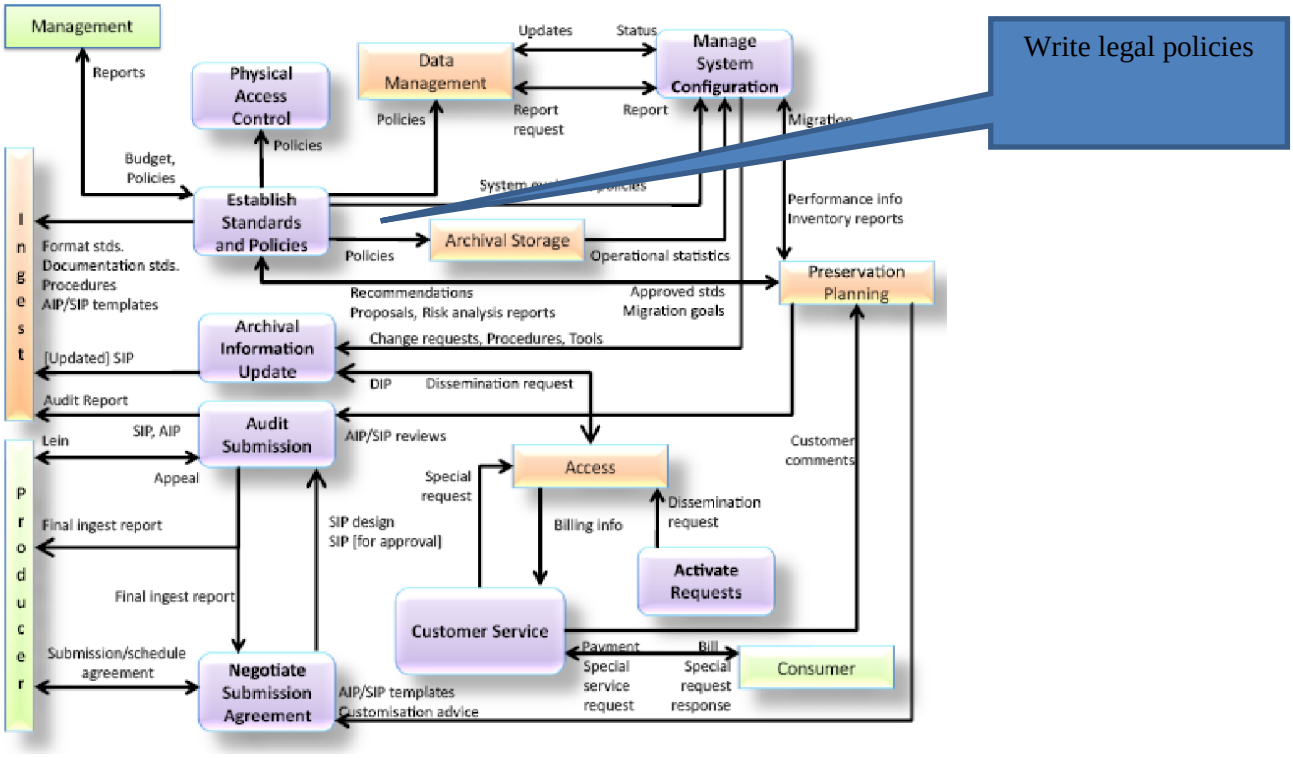


Figure 3.5: Functions of the Administration Functional Entity (Source: OAIS June 2012, page 4-11).

Function: Establish Standards And Policies

Action: Devise legal policies for access and security.

Description: This is an opportunity to refer to your organisation's existing legal policies on rights management, and apply them specifically to your blog repository. If such policies do not yet exist, this is an opportunity to create them. Notice how these policies can impact on the other OAIS functions.

See section 3.5.5 below for suggested legal policies, including definition of terms; IPR policy; notice and take down policy; and removal policies.

See section 3.5.4 for other policies, including policies for instructing the crawler and stopping captures.

Outcome / result: Your instance of BlogForever becomes integrated with your organisation's existing policies on rights management. Your blog-archiving activities will be legally sound. The legal policies will be published and available.

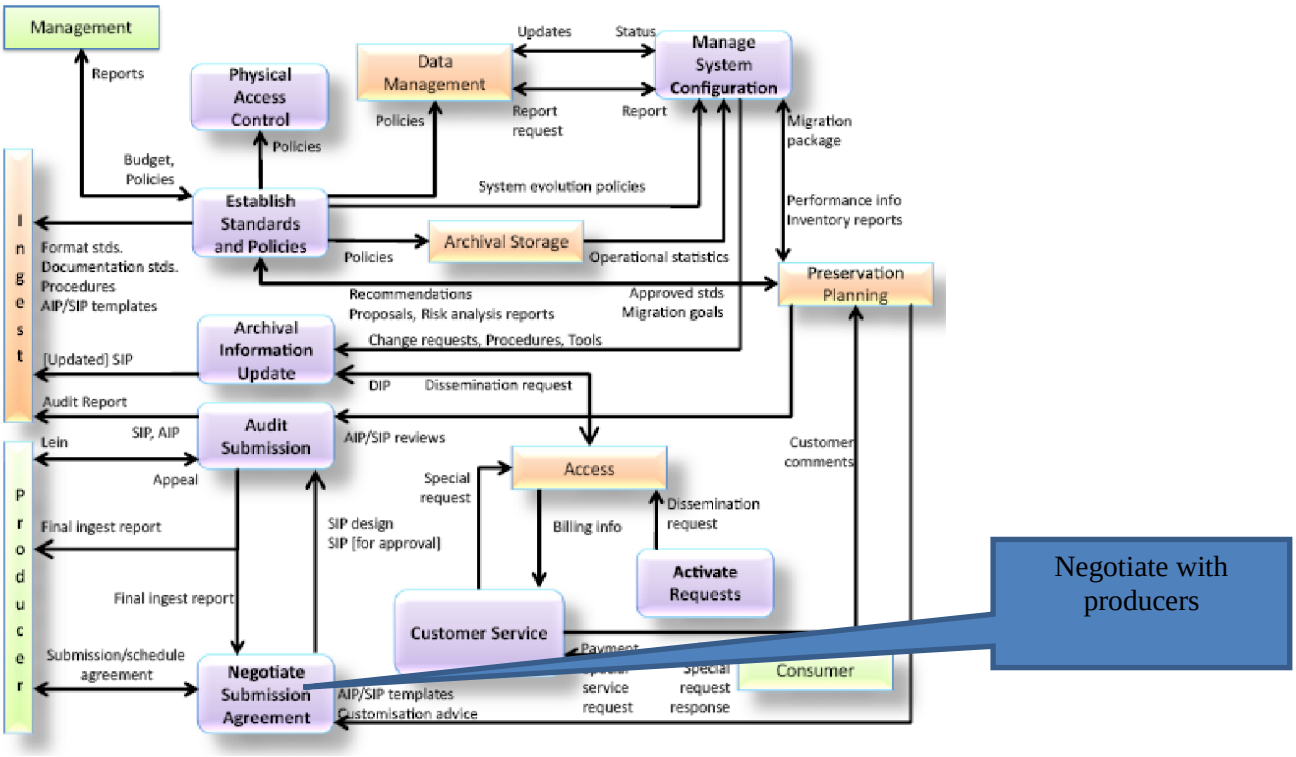


Figure 3.6: Functions of the Administration Functional Entity (Source: OAIS June 2012, page 4-11).

Function: Negotiate Submission Agreement

Action: Negotiate with producers

Description: This could be an agreement with a blogger or with a blog rights owner. You could require that bloggers to sign up to a deposit agreement before they deposit their blog with your repository, or secure an agreement with them before you proceed with a blog harvest. You can ask what types of access or usage are permitted by their license.

Outcome / result: When producers grant a copyright licence, they are permitting you to make a copy of the blog and to store it in an archive on your servers. They are also granting permission for you to take the necessary steps to preserve the blog, and to make it accessible to the public via the Internet now and in perpetuity.

See section 3.5.3 below for detailed suggestions on negotiating with producers.

The Negotiate Submission Agreement function has already been defined in some detail in D3.1. (p 130):

The repository needs to be sure that permission to preserve is confirmed. This is expressed as a submission agreement with the producer of the blog content. This requirement will clearly be influenced by the project deliverable 3.3 on rights management. The OAIS model depict this negotiation process as something that can be automated through a nexus of templates and SIP designs, but it still requires a coherent rights policy underpinning it.

The project's current thinking on rights management is that there is some scope for adopting a mechanism similar to the Creative Commons automated license. When submitting to the repository a new blog to be archived, the user or administrator could choose a specific license for it, from a list

of licenses, perhaps via a drop down menu. This list could be a knowledge base built up through usage, and kept as a database. This is one possible workflow point where a license could be assigned to the blog. Under that mechanism, based on the chosen license, access to the blog's content would be regulated accordingly.

Another scenario would be for users or administrators to submit blogs through a submission form. The repository administrator / manager(s) can verify the information before accepting the blog submission. In cases where the plan is to import a large number of blogs, then an automated submission process could be deployed. If a significant percentage of these submissions originate from the same source, that could allow assigning the same license for all the submissions.

3.2.4 Opportunities in the Preservation Planning functional entity

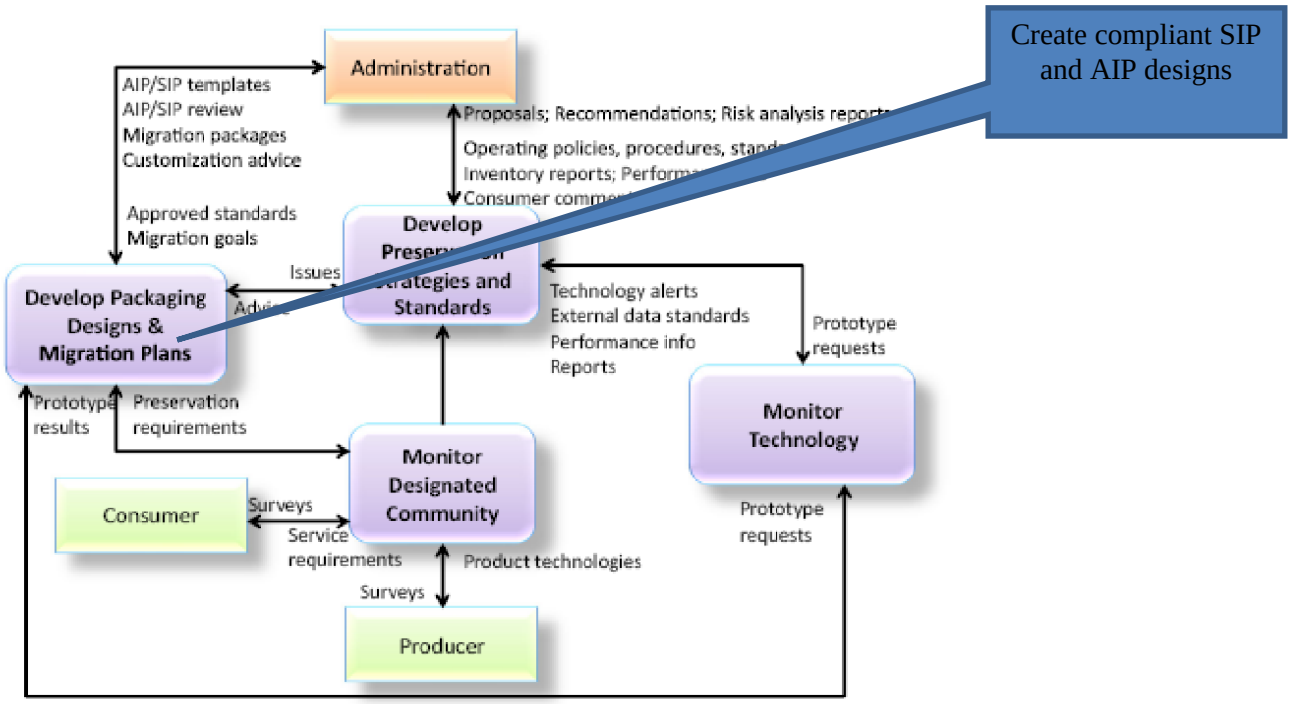


Figure 3.7: Functions of the Preservation Planning Functional Entity (OAIS June 2012, page 4-14).

Function: Developing Packaging Designs & Migration Plans

Action: Design legally compliant information packages.

Description: Here you can develop new "designs" for the information packages, and ensure that legal requirements are included in these designs. This planning function transmits new AIP and SIP designs to the Administration function and keeps them under review. It's an opportunity to incorporate legal requirements in terms of rights metadata and technical protection measures, and embed them directly in the information packages.

Outcome / result: The outcome ought to be Information Packages, customised to match your policies, that are legally sound.

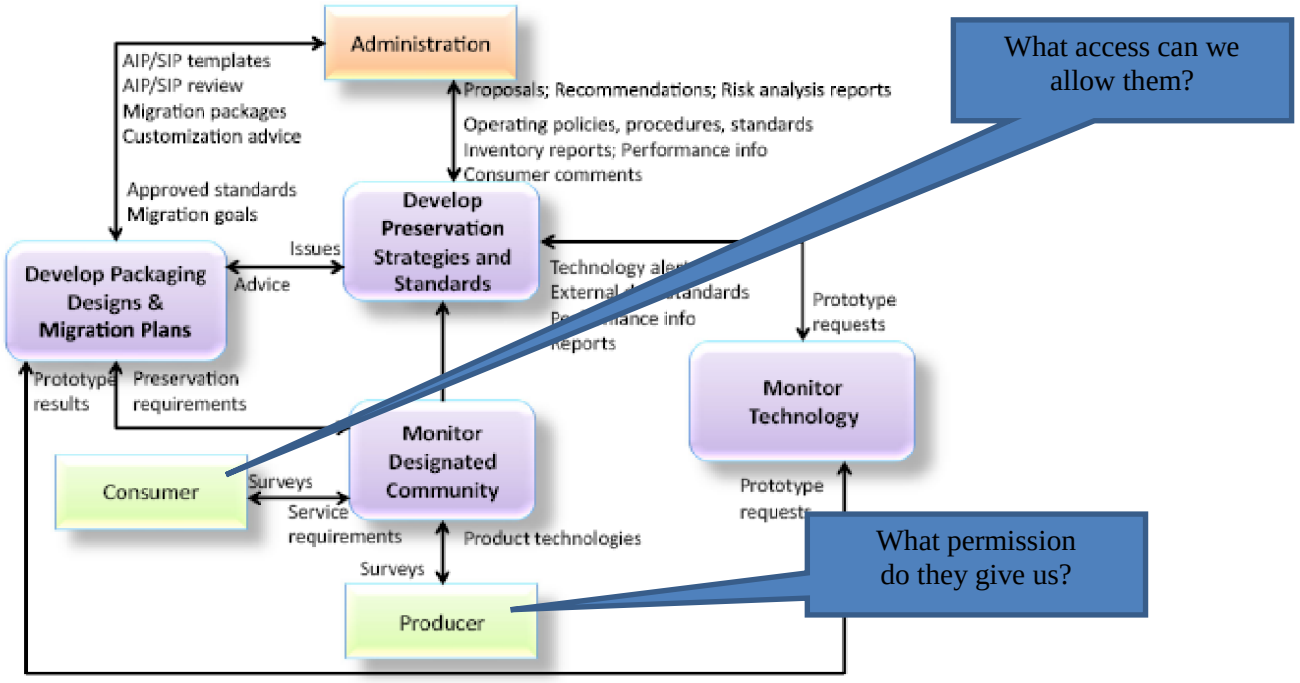


Figure 3.8: Functions of the Preservation Planning Functional Entity (OAIS June 2012, page 4-14).

Function: Monitor Designated Community

Action: Gather requirements from our Producers about rights.

Description: You can interpret this function as gathering requirements from your target bloggers about the rights they wish to assert, and what you can legally do with their content in the repository. The latter will allow you to inform your policy for the Consumers, advising them what they can do with the content.

Outcome / result: written policy statements and practices that will allow consumers to use the blog content, while protecting the rights of producers.

For detailed suggestions on managed access, see section 3.5.6 below

3.2.5 Opportunities in the Access functional entity

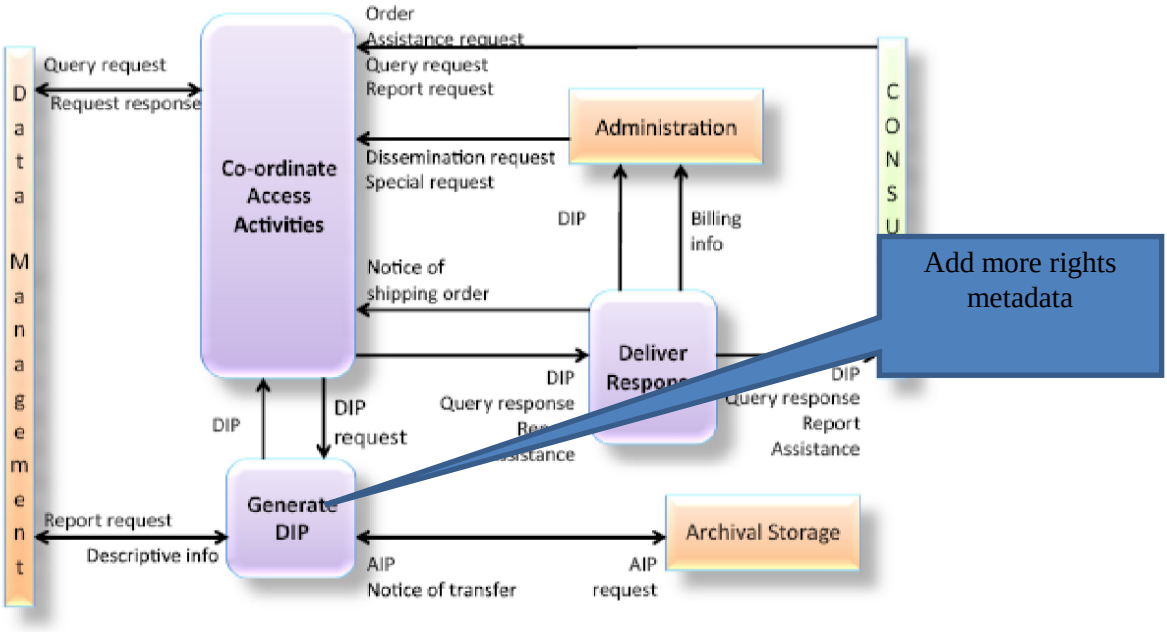


Figure 3.9: Functions of the Access Functional Entity (Source: OAIS June 2012, page 4-16).

Function: Generate DIP

Action: Add more rights metadata to the DIP.

Description: This is another opportunity to generate DIPs that have rights metadata embedded into the objects, or clearly attached to the blog record and declared as part of the public catalogue. If rights metadata has not been added to the original submission, this is another chance to do so.

Outcome / result: permanent records of rights information associated with the blogs you are harvesting in your organisation. Such records can be referred to in case of any disputes or issues. It is also possible to store the metadata in a database, thus enabling better management.

See section 5 on cataloguing rights metadata for further information, including examples and suggested cataloguing rules.

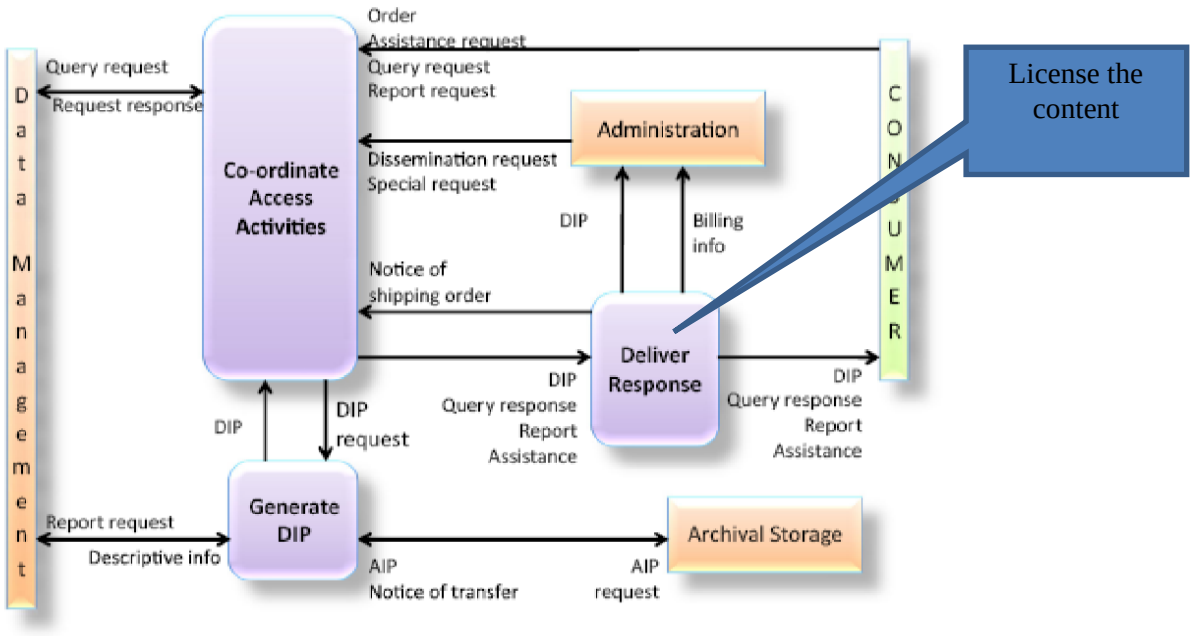


Figure 3.10: Functions of the Access Functional Entity (Source: OAIS June 2012, page 4-16).

Function: Deliver Response (1)

Action: License the content.

Description: This is the first of two possible interpretations of the OAIS function. As an enabling action, you can declare to the consumer what they can and cannot do with the archive blog. This will most likely take the form of a license, or a Creative Commons licenses. You might need to generate one license per blog, or there might be a single ruling policy that can apply to multiple blogs.

Outcome / result: documented / recorded evidence of agreements that will be essential in case of any future disputes.

For examples of a Permission agreement, see section 3.4.1 below; for a suggested consent form, see sections 3.4.2 and 3.4.3. See sections 3.4.4 – 3.4.6 for licensing, including Creative Commons.

Other outcomes are detailed in section 3.5.3, including the process of negotiating with rights owners, access restrictions, seeking copyright license, gaining permission from third-party rights holders, opt-out forms, and publishing statements about copyright implications.

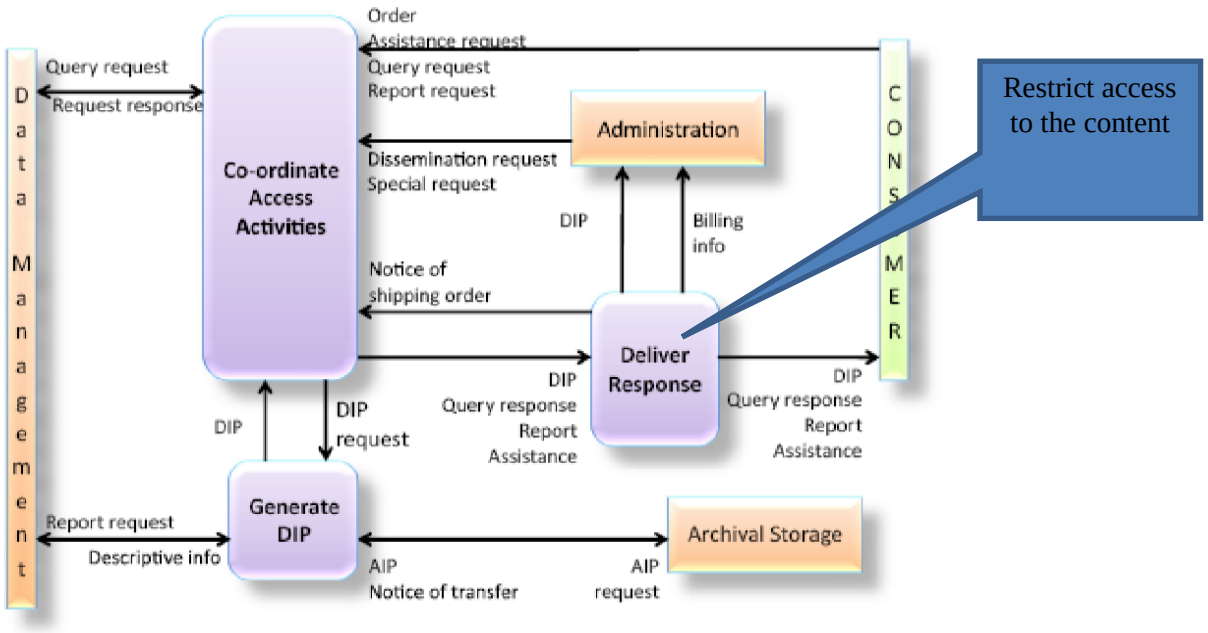


Figure 3.11: Functions of the Access Functional Entity (Source: OAIS June 2012, page 4-16).

Function: Deliver Response (2)

Action: Restrict access to DIPs.

Description: As a restrictive action, you can apply Technical Protection Measures at the point of access. This could be redaction, streaming content, using access copies of images with watermarks – anything that restricts or prevents copying or authorised use.

Outcome / result: This action ensures that sensitive blog data (if any) is never served or released to users, and thus protected from wrongful use at point of access.

3.3 Information Packages

For views of where the rights metadata sits, see section 6.1.4 of D3.1 on information packages.

Table 6.1-2 compares BlogForever with the OAIS terms for an Archival Information Package (AIP), and suggests that Rights Metadata will be part of what OAIS called Representation Information. The project's assumption at that time was that "Rights metadata will describe the rights associated with the blog" and that it "will be created by BlogForever administrators."

Table 6.1-3 compares BlogForever with the OAIS terms for a Dissemination Information Package (DIP), and suggests that Rights Metadata will be part of what OAIS called Representation Information. The project's understanding was that this would be rights metadata "for use by the consumers".

3.4 Possible actions related to the OAIS framework

We now discuss some possible actions and treatments that could take place in the OAIS workflow. This list is not comprehensive, and many other treatments are suggested in other chapters of this deliverable.

3.4.1 Permissions agreements

Declaration and Consent

Details of the Work

I hereby agree to deposit the following item in the digital repository maintained by Bangor University and/or in any other repository authorized for use by Bangor University.

Author Name: .....

Title: .....

Supervisor/Department: .....

Funding body (if any): .....

Qualification/Degree obtained: .....

This item is a product of my own research endeavours and is covered by the agreement below in which the item is referred to as "the Work". It is identical in content to that deposited in the Library, subject to point 4 below.

Non-exclusive Rights

Rights granted to the digital repository through this agreement are entirely non-exclusive. I am free to publish the Work in its present version or future versions elsewhere.

I agree that Bangor University may electronically store, copy or translate the Work to any approved medium or format for the purpose of future preservation and accessibility. Bangor University is not under any obligation to reproduce or display the Work in the same formats or resolutions in which it was originally deposited.

Bangor University Digital Repository

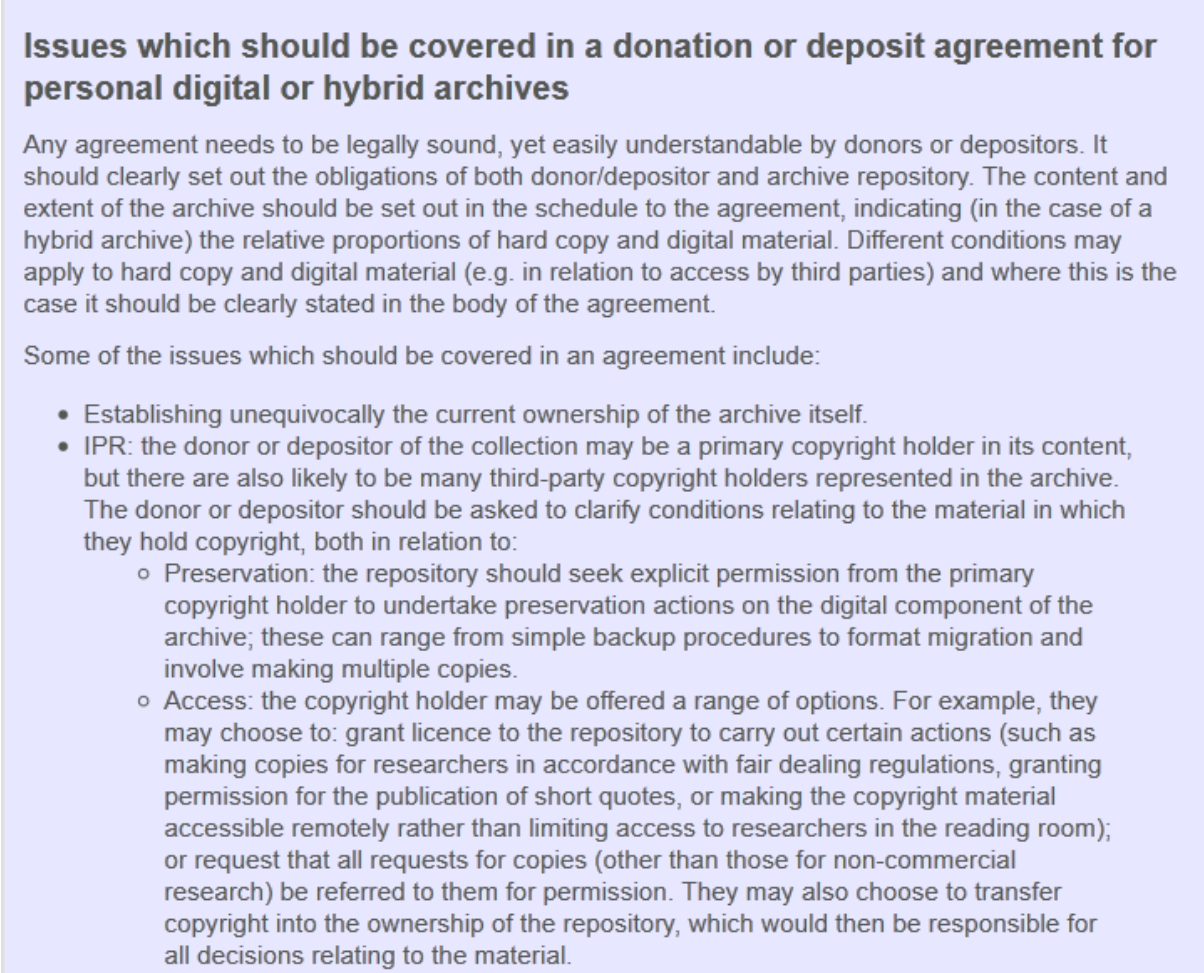
I understand that work deposited in the digital repository will be accessible to a wide variety of people and institutions, including automated agents and search engines via the World Wide Web.

Figure 3.12: The above example is a Permissions Agreement from Bangor University. You could create an adapted copy of this form that would allow bloggers to submit their blogs to your instance of BlogForever. Source: http://www.bangor.ac.uk/ar/main/publications/forms.php.en

In the above example Figure 3-12, notice:

- You are explicitly telling the Producer what you intend to do with their content.
- If the Producer completes this form of declaration and consent, then the Producers agree to let you make digital copies.
- You are also making sure they understand that a copy of the archived blog will be accessible on the web.

In this second example below Figure 3-13 from Paradigm UK, there is some advice about what to do when you are seeking explicit permission to preserve the blog:



Issues which should be covered in a donation or deposit agreement for personal digital or hybrid archives

Any agreement needs to be legally sound, yet easily understandable by donors or depositors. It should clearly set out the obligations of both donor/depositor and archive repository. The content and extent of the archive should be set out in the schedule to the agreement, indicating (in the case of a hybrid archive) the relative proportions of hard copy and digital material. Different conditions may apply to hard copy and digital material (e.g. in relation to access by third parties) and where this is the case it should be clearly stated in the body of the agreement.

Some of the issues which should be covered in an agreement include:

- Establishing unequivocally the current ownership of the archive itself.
- IPR: the donor or depositor of the collection may be a primary copyright holder in its content, but there are also likely to be many third-party copyright holders represented in the archive. The donor or depositor should be asked to clarify conditions relating to the material in which they hold copyright, both in relation to:
  - Preservation: the repository should seek explicit permission from the primary copyright holder to undertake preservation actions on the digital component of the archive; these can range from simple backup procedures to format migration and involve making multiple copies.
  - Access: the copyright holder may be offered a range of options. For example, they may choose to: grant licence to the repository to carry out certain actions (such as making copies for researchers in accordance with fair dealing regulations, granting permission for the publication of short quotes, or making the copyright material accessible remotely rather than limiting access to researchers in the reading room); or request that all requests for copies (other than those for non-commercial research) be referred to them for permission. They may also choose to transfer copyright into the ownership of the repository, which would then be responsible for all decisions relating to the material.

Figure 3.13: Outline of agreement issues, taken from the Paradigm UK workbook. Source: http://www.paradigm.ac.uk/workbook/record-creators/agreements.html

3.4.2 Consent Form

For BlogForever, many of your permissions needs could be met by a well-worded Consent Form which you can send to your blog owners to open the negotiation process.

If you can devise a standardised consent form, you can refer to it in your rights metadata by a simple citation. This would be a permanent record of what was agreed between you and the blogger at the time of archiving.

Issues to consider:

- Will the blogger allow you to crawl their blog?
- Who is the copyright holder?
- Are there other copyright holders besides the blog owner?
- What rights / restrictions etc. are they requesting?
- How do you express these in a license?
- Does copyright transfer to your repository?
- Can you republish their blog on a web-accessible platform?
- Will the blogger allow you to make digital copies for preservation?

- Can you transform the content and put into other accessible formats?

Your consent form can be tailored to meet the needs of your target bloggers. You need to ask questions about copyright, and about technical issues:

Copyright questions:

- Who is the owner of the content on the blog?
- Will you allow us to crawl your blog?
- Can we republish your content via our instance of the BlogForever platform?
- Are there third parties who need to be approached for consent?

Technical questions:

- Frequency of the crawl
- Depth of the crawl
- Turning off robots.txt
- Can we keep copies of the blog content on our servers?
- Can we perform preservation actions on these digital files?

For suggestions on how to enact the above, see the next section.

3.4.3 Elements of a Consent Form

This section is a digest of elements for rights management condensed from the publications of four web archives: PANDORA, the Web Archive of the National Library of Australia[79] ; The Library of Congress[80] ; Harvard's Web Archiving Collection Service (WAX)[81]; and The University of Michigan Web Archives at Bentley Historical Library[82].

The resources used were:

- Koerbin, Paul: Managing Web Archiving in Australia: A Case Study. 4th International Web Archiving Workshop (2004)
- PANDORA Web Archives: Services to Publishers [83]
- PANDORA Web Archives: NLA Copyright License FAQs [84]
- Library of Congress: Web Archiving FAQs [85]
- About WAX - Web Archive Collection Service - Harvard University Library [86]
- Deromedi and Shallcross: The University of Michigan Web Archives: Collection Development Policy and Methodology Version 1.1 (2011)

When drafting a form like this, it's unlikely that your form will need every single one of these elements. Once you have identified the elements that are suitable for your blog-archiving programme, express these as a written policy, and then recast them into a Consent Form for the blog owner when you seek consent. A list of FAQs could be appended to the Consent Form.

79 http://pandora.nla.gov.au/
80 http://www.loc.gov/webarchiving/
81 http://wax.lib.harvard.edu/collections/home.do
82 http://bentley.umich.edu/dchome/webarchives/
83 http://pandora.nla.gov.au/publishers.html
84 http://pandora.nla.gov.au/licencefaq.html
85 http://www.loc.gov/webarchiving/faq.html
86 http://wax.lib.harvard.edu/collections/about.do

We divide the elements into two main types: Copyright elements relating to ownership of blog content, and Technical elements relating to the crawl.

Copyright elements

Element	Action	Considerations
Identify owner of copyright	Find the copyright owner for the blog. This could be the original blogger, but it might also be a commercial publisher, a webmaster or the editor of a journal. For contributors to the blog, see next.	The copyright status of a blog *always* remains with the owner. The creator or publisher [blogger] retains copyright in both the original publication on the publisher's site, as well as in the copy in your instance of BlogForever.
Identify third party owners	Contributors to the blog may also hold copyright. Ask the blogger if they can assist with identifying contributors.	Pandora: "If others also own copyright in the publication, then they must also be agreeable to the granting of a copyright license. If you are not the sole copyright owner, we would appreciate it if you would gain the permission of the others before granting the copyright license. We assume that in granting permission all contributors to your publication are informed and in agreement that their work will be archived by the National Library."
Allow opt-out	Distribute communications to these content owners to explain the purpose of the blog archive, and inform them of their right to opt out. Opt-out could be managed with an online form.	LOC: "If you are a copyright owner of or otherwise have exclusive control over materials presently in the archive, you can opt out of online access to your site by completing this form."
Identify password-protected content	Do not archive password-protected content, unless by special permission from the blog owner.	
Gain permission to republish blog	Distinguish 'archived' sites from 'live' content with a prominent banner and statement at the top of each preserved web page.	See Michigan p 3
Identify access restrictions	Content owners may request that portions of their site be suppressed from public view.	See Michigan p 16
Gain permission to store copies on your servers	As part of BlogForever, you will be storing copies of the blog content on your own servers. Advise the blog owner of what you are doing and ensure they understand they are giving you permission to do this.	Pandora: "When you grant the National Library a copyright license, you are permitting it to make a copy of your publication as it appears on your Web site and to store it in an archive of Australian Web publications on the Library's own server."
Gain permission to preserve	As part of blog preservation in	Pandora: "When you grant the

	BlogForever, you may be making further copies of the blog content in the repository for preservation purposes, including migration / format shifting. Advise the blog owner of what you are doing and ensure they understand they are giving you permission to do this.	National Library a copyright license, you are permitting …the Library to take the necessary steps to preserve your publication…and to make it accessible to the public …in perpetuity."

Technical elements

Element	Action	Considerations
Frequency of crawl	State how often you intend to crawl the blog. This can also include start and end dates of the archiving action.	LOC: "Typically the Library crawls a website once a week or once monthly, depending on how frequently the content changes. Some sites are crawled more infrequently—just once or twice a year. The Library may crawl your site for a specific period of time or on an ongoing basis. This varies depending on the scope of a particular project. Some archiving activities are related to a time-sensitive event, such as before and immediately after a national election, or immediately following an event. Other archiving activities may be ongoing with no specified end date."
Depth of crawl	Indicate to the blogger which types of content will not be crawled.	LOC: "The Heritrix crawler is currently unable to archive streaming media, "deep web" or database content requiring user input, and content requiring payment or a subscription for access. In addition, there will always be some websites that take advantage of emerging or unusual technologies that the crawler cannot anticipate."
robots.txt	Notify the blogger about the crawl. Ask their preferences for prohibiting or allowing robots.txt. If allowed, then configure the spider to ignore robots.txt exclusions. If denied, then configure the spider to respect all exclusions in robots.txt files and do not capture any content designated as off-limits by the blogger.	See Michigan p 16. LOC: "The Library notifies site owners before crawling which means we generally ignore robots.txt exclusions." Harvard Wax: "You may specifically instruct our crawler to harvest material from your site or not to harvest material from your site by updating your robots.txt file to include us. The robots.txt file must be placed at the root of your server."

	Another possibility is asking the blogger to update their own robots.txt file appropriately.	
Stopping the crawl	Stop a capture if the spider detects any degradation of service or negative impact on the host's web server. Provide an online form so that blog owners can contact you immediately to stop a crawl.	See Michigan p 16

3.4.4 Licensing

Licensing is a way of managing copyright. It's about striking a balance between protecting rights, allowing people to use digitised content, and preventing exploitation (especially commercial) of the content.

In a future instance of BlogForever, general approach might involve: honouring the access requirements of the copyright owner, connecting to your organisation's existing rights policy and strategy, if any, enabling some use of the archived blogs in your collection, preventing commercial exploitation of the blogs by third-party publishers, making licenses clear and simple to understand, and enabling users to realise what risks might be involved at the point of download when they access a blog.

3.4.5 Creative Commons Licenses

Creative Commons [87] is a simple online methodology which may help your archivist / curator to devise licenses for your blog collection in BlogForever.

It may also prove helpful for enabling the bloggers to license their own content. Some blog submitters may already have applied Creative Commons to their own work, which may facilitate your task. Creative Commons have published advice and best practice [88] directed at bloggers for how to use Creative Commons, including a page for Google Bloggers [89]. There is also advice published by popular blogging platforms, such as Wordpress [90].

Creative Commons "develops, supports, and stewards legal and technical infrastructure that maximizes digital creativity, sharing, and innovation." It is intended to be an enabling process, to make the task of licensing online content very simple, without the need for legal advice. "Licensing a work is as simple as selecting which of the six licenses best meets your goals, and then marking your work in some way so that others know that you have chosen to release the work under the terms of that license."

Depending on the settings of the Creative Commons license, consumers / users of the content:

- Must always give attribution to the creator

87 http://creativecommons.org/
88 http://wiki.creativecommons.org/Blogger
89 http://wiki.creativecommons.org/Publish/Text/Blogger
90 http://en.support.wordpress.com/creative-commons/

- May or may not modify the content
- May or may not put the content to commercial use

By using the online Licensing Tool at Creative Commons, it is possible to generate a small block of code, which can also appear as an icon on a web page. This icon has the advantage of being recognisable by many users throughout the world; and it is machine-readable.

Creative Commons is very limited, and it isn't legally binding. However, it is a quick way of freeing up content without involving a lawyer.

3.4.6 Use Case for Creative Commons

Digital Bodleian [91] considered their strategy for their digitised content. In their case:

- They were "committed to providing clear and simple licensing for the use of … digitized content."
- Ideally, a single license would be applied across all of [their] publicly available digital collections, or they would at least minimize the exceptions.
- There are two over-arching questions at the core of the rights and licencing policy:
    - What rights will we assert over the digital copies of the public domain works in our collections?
    - In what conditions do they assert these rights?

NB: this case is not directly applicable to blog collections, as Digital Bodleian were seeking a way to make digitised content (i.e. scanned text and images from books) accessible. This is not directly comparable to born-digital content on blogs.

3.5 Rights management: best practices

This section is a digest of best practices for rights management condensed from the publications of four web archives: PANDORA, the Web Archive of the National Library of Australia; The Library of Congress; Harvard's Web Archiving Collection Service (WAX) ; and The University of Michigan Web Archives at Bentley Historical Library .

The resources used were:

- Koerbin, Paul: Managing Web Archiving in Australia: A Case Study. 4th International Web Archiving Workshop (2004)
- PANDORA Web Archives: Services to Publishers .
- PANDORA Web Archives: NLA Copyright License FAQs .
- Library of Congress: Web Archiving FAQs
- About WAX - Web Archive Collection Service - Harvard University Library
- Deromedi and Shallcross: The University of Michigan Web Archives: Collection Development Policy and Methodology Version 1.1 (2011)

Citations for these resources have already been stated in section 3.4.3.

Additional sources consulted on Notice and Take-Down policies were the Digital Public Library of America (DPLA), the British Library, The National Archives (UK), the Internet Memory Foundation, the European Archive, and the Internet Archive.

91 http://bdlssblog.bodleian.ox.ac.uk/archives/197

We would recommend that any curator who intends to create an instance of the BlogForever platform take advice from these and similar documents, and not just for rights management; they contain useful advice for many aspects of the entire web archiving lifecycle.

3.5.1 The OAIS workflow revisited

One section of this deliverable has described how a curator using BlogForever can think about responses to their legal issues in relation to the OAIS workflow. However the project brief for D3.3 correctly identified that there will be stages in the lifecycle *outside* those functional entities already identified in the OAIS framework. Such stages could include:

> **Creation**: A pre-ingest stage at which the target blogs are created by their bloggers. In most cases, this stage will probably not impact on your collections policy, unless you are in a position to influence your target bloggers. Intervention at the creation stage of the lifecycle is generally recommended as good practice for records managers, rather than archivists.
>
> **Selection**: A pre-ingest stage where your organisation can "evaluate [blogs] and select them for long-term curation and preservation, through adherence to documented guidance, policies or legal requirements". (Source: DCC Lifecycle Model [92])
>
> **Disposal**: A later stage where you can "dispose of [blog] data which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically such data may be transferred to another archive, repository, data centre or other custodian. In some instances data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction." (Source: DCC Lifecycle Model)

The stage that provides a large number of practical rights management opportunities to a blog archivist is **Selection**. In this document we will therefore identify many opportunities for rights management at the Selection stage – i.e. **before** the blogs are ingested into your instance of the BlogForever platform.

We will also identify other pertinent details from the above set of best practice documents, and thus provide further rights management opportunities by aligning them with the high-level functional entities in the OAIS framework.

This is a simple way of presenting possible rights management actions. It will be seen that the majority of possible rights management actions take place at Selection, Ingest, or Access.

Square brackets in the tables below indicate where we have interpreted the published best practice to show how it would apply to blog archives.

3.5.2 Scope of this section

The goal of this section is to provide guidance, advice and examples of best practice, and so enable anyone who builds an instance of BlogForever to devise suitable rights management policies and practices. This means that it may not be essential or desirable for you to try and implement all of the suggestions in these tables.

3.5.3 Selection stage

92 http://www.dcc.ac.uk/resources/curation-lifecycle-model

Definition: A pre-ingest stage where your organisation can "evaluate [blogs] and select them for long-term curation and preservation, through adherence to documented guidance, policies or legal requirements". (Source: DCC Lifecycle Model [93])

In this stage there are numerous opportunities to negotiate with blog owners and seek permission to crawl; identify copyright owners; create and draft licenses, or consent forms; and other actions to mitigate any risks associated with rights. In many cases it is important to seek permission before you start crawling a blog.

Action	**Description**	**Source**
Negotiate with rights owners	In Australia, archiving permission must be sought; and the permission status (i.e. granted, denied, unknown) is recorded in the management system.	Koerbin p13
Negotiate with rights owners	The scalability of the selective approach, which allows for the *negotiation of permission to archive with rights owners* and quality assurance in the archiving process, supports the building of an accessible, functional and undoubtedly valuable web archive.	Koerbin p 23
Negotiate with rights owners	Inform individual content owners of their rights.	Michigan p 3
Negotiate access restrictions	As part of the process of obtaining permission to archive it may be necessary to negotiate access restrictions (typically in the case of commercial publications).	Koerbin p 13
Seek a copyright license from creators and publishers	In Australia, because the legal deposit provisions of the Commonwealth Copyright Act 1968 do not include electronic publications, the Library and its partners must seek a copyright license from creators and publishers before it can copy a publication [blog] into the Archive.	PANDORA (Services to Publishers)
Gain permission from third-party rights owners	If others also own copyright in the publication [blog], then they must also be agreeable to the granting of a copyright license.	PANDORA (Services to Publishers)
Create and manage licenses	In Australia, when you grant the National Library a copyright license, you are permitting it to make a copy of your publication as it appears on your Web site [blog] and to store it in an archive of Australian Web publications	PANDORA (NLA Copyright License FAQs)

93 http://www.dcc.ac.uk/resources/curation-lifecycle-model

	on the Library's own server. You are also permitting the Library to take the necessary steps to preserve your publication, and to make it accessible to the public via the Internet now and in perpetuity.	
Provide an opt-out form	In the USA, if you [the blogger] are a copyright owner of or otherwise have exclusive control over materials presently in the LOC archive, you can opt out of online access to your site by completing a form.	LOC
Provide an opt-out form	UARP will distribute communications to content owners to explain the purpose of the University of Michigan Web Archives, inform them of their right to opt out or suppress content, and invite questions or concerns.	Michigan p 16
State the copyright implications of blog archiving	It is important to note that the creator or publisher retains copyright in both the original publication on the publisher's [blog], *as well as in the copy in PANDORA.*	PANDORA (Services to Publishers)
State the copyright implications of blog archiving	In the USA, the copyright status of your site [blog] remains with you [blog owner].	LOC
Use a permissions tool	The Library of Congress uses a permissions tool that allows easy contact with [blog] owners via e-mail, and enables the [blog] owners to respond to permissions requests using a web form. The responses are then recorded in a database.	LOC
Identify your crawler to the blogger in advance	The Harvard WAX crawler (hul-wax) is identified on their public website.	Harvard

3.5.4 Ingest stage

Definition: The ingest stage (or functional entity) is the point at which blog content is crawled and accessioned into your instance of the BlogForever repository.

In OAIS terms, it "contains the services and functions that accept Submission Information Packages from Producers, prepares Archival Information Packages for storage, and ensures that Archival Information Packages and their supporting Descriptive Information become established within the OAIS."

In this stage there are opportunities for authoring basic rights metadata as part of the ingested package. There are also opportunities for programming the spider appropriately, depending on how much permission you have.

Action	Description	Source
Add copyright notes to metadata	In PANDORA, rights management metadata includes publisher details, permission status, registry file reference, and access restrictions.	Koerbin
Instruct the crawler	The Library of Congress notifies site owners before crawling which means we generally ignore robots.txt exclusions.	LOC
Instruct the crawler	The Harvard WAX crawler will obey all common instructions in robots.txt files. You [blogger] may specifically instruct our crawler to harvest material from your site or not to harvest material from your site by updating your robots.txt file to include us. The robots.txt file must be placed at the root of your server.	Harvard
Instruct the crawler	The WAS web crawler is configured to respect all exclusions in robots.txt files and will not capture content designated as off-limits by a webmaster.	Michigan p 16
Stop captures	The Library of Congress always tries to politely crawl sites in order to minimize server impact.	LOC
Stop captures	WAS will stop a capture if it detects any degradation of service or negative impact on the host's web server.	Michigan p 16

3.5.5 Planning stage

Definition: At this planning stage, your organisation has opportunities to create policies regarding rights management. We propose three policies which will assist with rights management: defining the terms of a license; a written policy about how you will respect the intellectual property rights of blog owners, and a notice and take-down policy, enabling the removal of repository content which is in breach of copyright.

In OAIS terms, this equates to the "Preservation Planning Functional Entity", which "provides the services and functions for monitoring the environment of the OAIS and which provides recommendations and preservation plans." It has explicit responsibility for creation of repository policies.

Action	Description	Source
Define terms	In Australia, you still retain full copyright in your publication,	PANDORA (NLA Copyright License FAQs)

	both in the original version on your Web site [blog] and the archived version in the National Library's archive. What am I doing when I grant the National Library of Australia a copyright license? You are granting the Library limited rights to copy your publication and provide access to it, in perpetuity.	
Write and publish a policy respecting IPR	In Michigan, the University Archives and Records Program (UARP) will respect the intellectual property rights of content owners. UARP strives to respect the rights of content owners and to follow professional best practices for intellectual property rights management in website [blog] preservation.	Michigan pp 3, 16
Notice and Take Down policy	This is a very useful means of minimising rights issues arising from republishing archived blogs. Once the policy is written, it can be enacted by creating an online submissions form for blog content owners.	
Notice and Take Down policy	Harvard's form of words: "If you own or control copyrighted content available in WAX and wish it to be taken down, please let us know. To make a takedown request or inquire about inclusion of your content in WAX, go to Questions and Comments. Please identify in your submission the URL(s) of the web page(s) carrying your content, the date(s) and time(s) of archiving, the specific content on the page(s) to which you claim rights, and the nature of your rights."	Harvard
Removal policy	The Removal policy should describe clearly the conditions under which content can and will be removed from the repository, including instructions for the submission of complaints, a description of how complaints are handled and any other circumstances under which content may be removed without the submission of a complaint (for example, if the archive discovers illegal content in the repository, or in the case of robots.txt).	

Removal policy	Editing or deleting your User Content will alter the public availability of the User Content, but may not permanently delete the content from the Services.	DPLA
Removal policy	The [Library] may, in appropriate circumstances and at its discretion, remove certain content or disable access to content that appears to infringe the copyright or other intellectual property rights of others.	British Library
Removal policy	To remove your site from the Wayback Machine, place a robots.txt file at the top level of your site (e.g. www.yourdomain.com/robots.txt) . If you cannot put a robots.txt file up, read our exclusion policy. If you think it applies to you, send a request to us at info@archive.org.	Internet Archive http://archive.org/about/terms.php
Removal policy	If you like to view, correct, complete or remove your personal information, please contact the Archive at info@europarchive.org	European Archive
Removal policy	Material will be taken down temporarily on receipt of a request from a member of the public or a government department. The case will then be considered by a Takedown Panel composed of members of staff who provide relevant expertise.	The National Archives (UK)
Removal policy	Content that is known to breach the law will not be included and access will be removed in respect to content that is subsequently proven to be in contravention of the law.	Australian National Archives, PANDORA

3.5.6 Access stage

Definition: this is the point at which you make archived blog content available to your user community through the BlogForever platform. There are numerous strategies and opportunities allowing you to manage this access.

In OAIS terms, the Access functional entity "contains the services and functions which make the archival information holdings and related services visible to Consumers".

Action	**Description**	**Source**

Action	Description	Source
Link to publisher's copyright statement	On the front end of BlogForever, publish a link to the archived version of the publisher's copyright statement. (This is one of several display options.)	Koerbin p 19
Restrict access at title level	In PANDORA, access restrictions are set at the title level. Three types of access restriction can be applied: period restriction, date restriction and authenticated restriction.	Koerbin p 13
Define access by location	In Australia, period and date restrictions are applied in conjunction with locations to which the access is limited. For example, access may be restricted to staff-only areas of the National Library or to a single PC in the Library's Main Reading Room.	Koerbin p 14
Define access by location	In the USA, if you [blog owner] deny off-site access, the Library may catalog and identify the [blog] as part of a particular collection on our public website, but your archived [blog] will only be available to researchers who visit the Library of Congress buildings in Washington, D.C. and by special arrangement	LOC
Manage copyright and access restrictions via the license	In the case of Web publications that have been archived by the Library, the catalogue record contains an active link both to the version of the publication on the publisher's site and to the version in the Library's Archive. When a researcher opts to look at the version in the Archive, a title entry page displays first, providing information about the title, including a link to the publisher's site. A general statement about copyright is included on the title entry page and a link to the publisher's own copyright statement is also provided. If access to a publication needs to be restricted for some reason (for instance, it is a commercial title, or the contents are culturally sensitive) the Library can ensure that necessary protective measures	PANDORA (NLA Copyright License FAQs)

Action	Description	Source
	are put in place.	
Defer publishing	In the USA, if you [blog owner] decide to allow the Library to provide online access to your archived [blog] to researchers, the Library will not provide access until at least a year after the web archiving.	LOC
Defer publishing	Embargo archived content for six months after capture so that the archived copy will not be mistaken for the original or divert viewers from the 'live' site.	Michigan p 3
Publish copyright statement at collection level	LOC have a statement on each collection homepage about copyright.	LOC
Distinguish live blog from archived blog	In LOC, there will be a banner at the top of the page that alerts researchers that they are viewing an archived version.	LOC
Distinguish live blog from archived blog	Distinguish 'archived' sites from 'live' content with a prominent banner and statement at the top of each preserved web page [blog].	Michigan p 3, 16
Restrict access on request	Suppress [blog] content from public view or refrain from website [blog] preservation at the request of content owners. Content owners may request that portions of their [blog] be suppressed from public view and can choose to opt out entirely from captures.	Michigan p 16
Publish a disclaimer	Having taken such steps as are reasonable, publish a disclaimer saying that best effort been made to identify the copyright owners, apologizing for any infringement and inviting copyright owners to make contact.	

3.6 Conclusions

Given the level of detail provided in section 3, we are aware that implementing rights management can seem like a daunting task. There's a lot of information, standards, best practices and guidelines. To make rights management more achievable, here we provide a summary list of things to consider when adhering to best practice:

- **Ask yourself what is appropriate for your collection**. To do this, understand the scope of the blogs that you intend to capture as part of the initial selection process. Do you know who the copyright owners are? Can you get their permission easily? If the collection is small and well-defined, the rights situation could be much easier than you think (e.g. all academics automatically consent to have their blogs copied as a condition of their job).

- **Ask what is applicable to your organisation**. What are the risks for reputational damage, financial loss and embarrassment to your organisation if you breach copyright laws?
- **Always ask permission from a blog owner**, and keep copies of your emails. Even if they ignore you, there is still a record that you asked them.
- **Automate the permission-seeking process** as much as possible. Even using a MailMerge can save you a lot of time.
- **Learn from others**. All web archivists have faced these (or similar) rights management problems, so you may not have to reinvent everything. You can also learn from experience of digital librarians, archivists, and other information professionals, regarding the IPR and copyright of digital content (other than blogs).
- **Create and keep records**. Keep all of your user agreements, license, copyright statements etc. in case of disputes. Where possible, express what you can within your rights management metadata. Keep copies of all your email negotiations with blog owners. Good records will show that you have made every effort to remain compliant with copyright law.
- **Make policies for your blog-archiving activities**. Write them down, keep them up to date, and make sure they are understood within the organisation. Where they affect your users and bloggers, publish and disseminate these policies wherever and whenever possible.
- **Create forms**. Create forms for consent, permission, and licensing, and use them.
- **Keep communicating**, to users and bloggers, what you intend to do with your blog archive; increase their trust and your credibility.
- **Build a workflow** for the rights management process. We have suggested OAIS. This might not be best for you. We have provided it in this report as one example. If you are not actively seeking to be compliant with that standard, then think again.
- **Let people opt out** of your blog-archiving process.
- **Take down any infringing content immediately**.
- **Do what is achievable**. Don't let rights management concerns stop your project dead. If need be, take a calculated risk that will allow you to archive at least a proportion of the blogs in your target collection.
- Remember that all of rights management is a form of **risk management**. All you can do is mitigate the risks, or in some cases avoid them.
- Some will tell you that technically speaking, all web-archiving (or anything that involves making digital copies) is illegal under copyright laws. That may be true, but **it does not allow you to ignore the law**.

4 BlogForever Repository and Spider Functionalities that support DRM

This section is intended to present an overview of the capabilities of the BlogForever Spider and Repository in relation to rights management. Given rights-related data, and points at which we might catalogue rights metadata (more on metadata will be discussed in Section 5), the objective in this section to explore how we might retrieve the necessary information to be catalogued.

More specifically, it describes:

1. what information the BlogForever spider currently harvests about rights associated with blogs (intellectual property, licenses, privacy, and policies);
2. how the rights information collection process is implemented;
3. what rights information is passed to the BlogForever repository;
4. how the rights information collected by the BlogForever spider is processed by the BlogForever repository to support digital rights management;
5. questions that might arise about the BlogForever repository capabilities;
6. an analysis of the content harvested by the BlogForever spider, to diagnose the potential for obtaining rights metadata automatically.

In this section, we will refer to any document and/or information retrieved by the spider for this purpose as "policy data".

4.1 The BlogForever Spider and Policy Data

The BlogForever spider incorporates an automated process for collecting links to some policy documents. This is implemented through a search for links where the hypertext associated with the link contains selected terms associated with rights and policies. There are five terms being used currently. These are:

- privacy policy
- terms of use
- term of use
- terms of service
- copyright

15.xml	wlwmanifest	BlogData	OK	1,03 KB	http://cnnwhatsnext.wordpress.com/wp-includes/wlwmanifest.xml	text/xml	Preview
16.html	shortlink	BlogData	OK	164,35 KB	http://wp.me/1njmc	text/html; charset=UTF-8	Preview
17.html	openid.server	BlogData	OK	24 Bytes	http://cnnwhatsnext.wordpress.com/?openidserver=1	text/html; charset=utf-8	Preview
18.html	openid.delegate	BlogData	OK	164,35 KB	http://cnnwhatsnext.wordpress.com/	text/html; charset=UTF-8	Preview
19.xml	What's Next	BlogData	OK	1,52 KB	http://whatsnext.blogs.cnn.com/osd.xml	application/xml	Preview
20.xml	WordPress.com	BlogData	OK	1015 Bytes	http://wordpress.com/opensearch.xml	text/xml	Preview
21.html	Terms of service	Privacy Policy	OK	48,95 KB	http://www.cnn.com/interactive_legal.html	text/html	Preview
22.plain		Robot	OK	772 Bytes	http://whatsnext.blogs.cnn.com/robots.txt	text/plain; charset=utf-8	Preview
23.jpeg		BlogSnapShot		1,49 MB	http://whatsnext.blogs.cnn.com/	image/jpeg	Preview
24.jpeg		BlogSnapShotThumbnail		22,73 KB	http://whatsnext.blogs.cnn.com/	image/jpeg	Preview

Figure 4.1: File capture results as provided by the spider. The item in the red box (tagged with the hypertext keyword "Terms of Service") indicates an instance of capturing a policy document.

The hypertext, link associated with the hypertext, and the documents associated with the link are all collected and passed on to the repository. The information transmitted in addition to the target document is presented in Figure 4.1.

The retrieval method applied to the BlogForever WP5 use case data, available at the time the work presented here was carried out, yields only 18 out of 1000 blogs that have any documents of the targeted type. This, however, could be a result of searching with a limited number of terms (no expansion using similar terms) only in English.

There were 30 documents altogether collected from the 18 blogs in question. An extension to this approach which uses terms across several languages and term expansion (e.g. using rights expression languages and metadata terms – see Section 4.3) could improve recall. This, however, may reduce precision, and may result in the requirement for post-processing to sort out the true policy data. Depending on the extent of language processing and information retrieval tools involved, this could become substantially complex, leading to a requirement for increased number of servers or alternative processing power.

Before venturing into such modes of increased complexity, which may be a full length report in itself, it is the purpose of the current report to present a scoping study of the potential for automating rights metadata extraction, based on the limited number of terms, languages, and blogs that the BlogForever spider has already been employing thus far.

4.2 The BlogForever Repository and Rights Management

The objective of this section is to elaborate on the functionalities of the BlogForever repository that support assigning values to metadata fields related to digital rights (Section 4.2.1), and, that might be frequently questioned by a future curator of blogs using the tool (Section 4.2.2). More technical detail on the architectures and functionalities of the BlogForever repository is available in deliverable BlogForever: D4.8 Final BlogForever Platform (August 2013).

4.2.1 Rights metadata

The BlogForever repository does not currently apply any automated method, using the data discussed in Section 4.1, for assigning rights information to metadata fields. The policy data received are linked to the blogs in question but its transformation into information that can be catalogued as rights metadata is not rigidly implemented. This allows flexibility for future blog curators to configure the software according to their needs.

There are, however basic rights metadata fields being used. This is derived from the model prescribed in the deliverable BlogForever: D2.2 Weblog Data Model (2011), and, is coded in MARCXML. These fields are (the MARC XML code is presented in the parenthesis):

- copyright (542)
- ownership rights (542)
- distribution rights (542)
- access rights (542)
- license (542 $f)

Examples of how these metadata can be catalogued is further discussed in Section 5.

As mentioned earlier, there is no automated method applied to the policy documents retrieved by the spider to extract rights metadata. However, a simple automated scenario of filling the fields would be to provide the link to any policy and agreements found within the blogs to all of these elements and alert the end-user and/or end-curator of their existence to post-process document for verification and further information. It is important that these fields can be edited easily by the repository manager without complicated processes involving the administrator. This is further discussed in the next section.

4.2.2 BlogForever repository capabilities related to DRM

In order to present the BlogForever platform implementation of rights management, we have formulated a set of frequently asked questions which outline all key related functionality. This is intended to clarify the flexibility of the software to be configured to meet the needs of the repository, for example, in relation to the best practices outlined in Section 3 and Section 5 for digital rights management and metadata cataloguing.

Frequently Asked Questions:

1. **Is there any metadata schema currently being used to record rights metadata in the repository** (for example, PREMIS[94] Rights in a METS[95] record)?

The metadata information from the spider is stored in MARC. A practical approach to storing this in METS has not yet been implemented. Tags to use and where to put them in METS need to be defined to meet organisational requirements. Generally speaking it is suggested that minimally the following information should be provided for each object[96]:

1. status
2. jurisdiction
3. note for naming rights holder
4. note for intended use
5. note for limits on use

Ideally, the law/act that applies, restrictions, and dates (start and end) should also be provided. Without explicit negotiation with the copyright holder (that is, material being collected by web crawling technology), it is unclear whether even the minimal information can easily be confirmed.
Content collected automatically comprise vague statements or simply links to the terms, but never a machine readable sentence saying what the rights associated to the blog are. We can collect information from the submitter when they submit the URL of the blog to be harvested. But the submitter is not necessary the copyright holder and the accuracy of the submitted information cannot be verified at the time of submission.

2. **Is the curator able to directly assign or modify rights metadata to items? Is it only the systems administrator who can do this?**

The software is capable of allowing/disallowing users or groups of users to perform tasks in the repository. Therefore, the administrators of the repository will decide if they authorise this action only for themselves, curators, any registered user, a group of trusted users, anyone, or whatever combination they may come up with.

3. **Is there already an agreed way of expressing the restrictions to the end-user? If so, what is the agreement?**

The BlogForever platform is fully capable of defining what restrictions and expression method to use in communicating right metadata. At the moment there is no fixed approach to doing this integrated into the system. A curator will need to think about which restrictions apply in their case and how to express them to the end user.

4. **What kind of mechanisms are in place (or planned to be in place for the final implementation) to control access based on the rights metadata?**

At the moment the BlogForever repository defines 3 visibility levels to objects. These are:

94 http://www.loc.gov/standards/premis/
95 http://www.loc.gov/standards/mets/
96 http://www.slideshare.net/Jacknickelson/premis-rights-implementation-at-university-of-california-san

1. Public: any visitor can see it
2. Restricted: any registered user can see it, but not the guests
3. Private: only the submitter and the administrators can see it

However, whoever deploys the software will decide which visibility option to apply to which content. The software is flexible enough that they can also introduce extra options and, for example, allow only a selected group of users to access the content.

It might be recommended that any information that is provided by the submitter of the URL should come with either 1) an agreement that they are responsible for the accuracy of the information provided, and/or, 2) confirmation that they are the rights holder of the information associated with the submitted URL or a representative thereof. For example, the use of a consent form was discussed in Section 3. In addition, it is recommended that the repository make provisions for terms of service to specify these conditions.

The question of expressing rights metadata will be revisited in Section 5, where we provide examples of accepted standards and cataloguing approaches. In addition, in the remainder of this section, we will see that, a basic level of text processing could provide a better indication of what information might be associated with each metadata field. This could be thought of as a compromise between leaving the end-user to take responsibility and search for the target information and finding and providing the precise information at the time of ingestion.

4.3 Text analysis of policy data

In this section, we present a brief analysis of the policy data retrieved by the BlogForever spider to explore the content, assess the potential of automatically extracting rights metadata, and/or, suggest ways of accessing information that might help DRM. The data for this study has, so far, been very limited. Therefore, the conclusions in this section cannot be definitive.

4.3.1 The data

As mentioned in Section 4.1, there were only 30 documents belonging to 18 blogs that were available for study. Later it was found that, due to a minor bug in the data extraction process (which was unfortunately not discovered in time for this study), not all the data was reliable for the study proposed here. As a result, we have only included an analysis of nine blogs (one hosted by the Financial Times[97] and eight hosted by the Guardian[98]), resulting in an examination of 18 documents. The blog URLs, hypertext, and document URLs have been provided in Table 4.1.

Blog URL	**Hypertext**	**Document URL**
blogs.ft.com/westminster/	Copyright	http://help.ft.com/tools-services/copyright-policy/
blogs.ft.com/westminster/	Privacy policy	http://help.ft.com/tools-services/financial-times-privacy-policy/
www.guardian.co.uk/law/baby-barista-blog/	Terms of service[99]	http://www.theguardian.com/help/terms-of-service
www.guardian.co.uk/law/baby-barista-blog/	Privacy policy	http://www.theguardian.com/help/privacy-policy

97 http://www.ft.com/
98 http://www.theguardian.com/
99 This hypertext was found to be “Terms & conditions” when examined in July 2013. The file was checked to see that it is the same document.

Blog URL	Hypertext	Document URL
www.guardian.co.uk/sport/blog/	Terms of service	http://www.theguardian.com/help/terms-of-service
www.guardian.co.uk/sport/blog/	Privacy policy	http://www.theguardian.com/help/privacy-policy
www.guardian.co.uk/stage/theatreblog/	Terms of service	http://www.theguardian.com/help/terms-of-service
www.guardian.co.uk/stage/theatreblog/	Privacy policy	http://www.theguardian.com/help/privacy-policy
www.guardian.co.uk/lifeandstyle/allotment/	Terms of service	http://www.theguardian.com/help/terms-of-service
www.guardian.co.uk/lifeandstyle/allotment/	Privacy policy	http://www.theguardian.com/help/privacy-policy
www.guardian.co.uk/science/the-lay-scientist/	Terms of service	http://www.theguardian.com/help/terms-of-service
www.guardian.co.uk/science/the-lay-scientist/	Privacy policy	http://www.theguardian.com/help/privacy-policy
www.guardian.co.uk/music/tomserviceblog/	Terms of service	http://www.theguardian.com/help/terms-of-service
www.guardian.co.uk/music/tomserviceblog/	Privacy policy	http://www.theguardian.com/help/privacy-policy
www.guardian.co.uk/uk/davehillblog/	Terms of service	http://www.theguardian.com/help/terms-of-service
www.guardian.co.uk/uk/davehillblog/	Privacy policy	http://www.theguardian.com/help/privacy-policy
www.guardian.co.uk/technology/gamesblog/	Terms of service	http://www.theguardian.com/help/terms-of-service
www.guardian.co.uk/technology/gamesblog/	Privacy policy	http://www.theguardian.com/help/privacy-policy

Table 4.1: Policy data: blog URL, hypertext, and document location.

The document URLs in Table 4.1 show that the "privacy policy" and "terms of service" documents of blogs hosted by the Guardian are located at the same URLs respectively. In fact, a test using Python's filecmp[100] shows that all the privacy policy documents from the Guardian are identical, as are all the terms of service documents. This conforms to what we might expect: this constitutes initial evidence that policy data on blogs hosted by the same organisation are shared across the blogs owned by the organisation[101]. The documents from the Financial Times blogs were not identical to any of the other documents.

In the following sections, we will compare the documents on the basis of key phrase frequency, document similarity, and hypertext link usage to see if there is a distinctive pattern that distinguishes or connects the four documents. We will carry out the analysis with respect to the four documents appearing in Table 4.2. The ID provided in the first column will be used in the figures presented.

ID	Blog URL	Hypertext	Document URL

100 http://docs.python.org/2/library/filecmp.html

101 Note that this is not a statement about the blogging platform provider, but the content providing/publishing organisation.

ID	Blog URL	Hypertext	Document URL
0	blogs.ft.com/westminster/	Copyright	http://help.ft.com/tools-services/copyright-policy/
9	blogs.ft.com/westminster/	Privacy policy	http://help.ft.com/tools-services/financial-times-privacy-policy/
10	www.guardian.co.uk/law/baby-barista-blog/	Terms of service[102]	http://www.theguardian.com/help/terms-of-service
11	www.guardian.co.uk/law/baby-barista-blog/	Privacy policy	http://www.theguardian.com/help/privacy-policy

Table 4.2: Documents being analysed in the text analysis.

4.3.2 Frequency Analysis

In this section we provide a comparison of the four documents in Table 4.2 of Section 4.3.1 with respect to frequency of the general terms that we introduced in Section 2 ("intellectual property", "copyright", "trademark", "patent", "license", "trade secret", "privacy", "statute", "defamation", "illegal content"). The result is displayed in Figure 4.2 and results show a higher frequency of the term *privacy* in the two privacy documents and the high frequency of the term *copyright* in the Financial Times copyright document. The terms of service document from the Guardian seems to use the word *content* more than any other document.

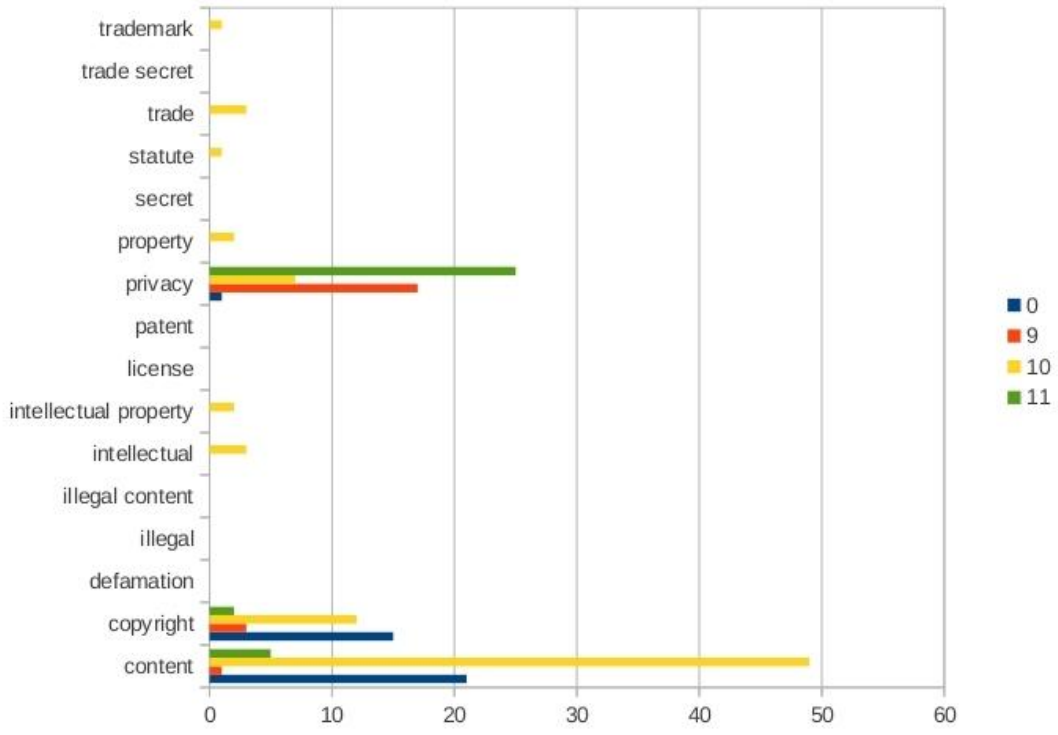


Figure 4.2: Frequency of general digital rights terms across the Financial Times copyright document (0) and privacy policy (9), and the Guardian terms of service document (10) and privacy policy (11).

102 This hypertext was found to be "Terms & conditions" when examined in July 2013. The file was checked to see that it is the same document.

We also used the vocabulary from Open Digital Rights Language (ODRL)[103] to carry out a similar analysis. This allows us to make the first steps in assessing the viability of a future curator using any one of these standards to express and manage digital rights. We have taken the head phrases (termed as *identifier* in the ODRL vocabulary[104]) to examine the frequency of the words in the phrases with respect to each document.

The vocabularies from Creative Commons Rights Expression Language (ccREL)[105], Dublin Core[106], and PREMIS[107] was also considered for analysis, but it was decided that ODRL covers the core elements of the other schemas. In fact, it also reflects a wide range of metadata that have been developed over the years such as:

- [RFC-2119] Key words for use in RFCs to Indicate Requirement Levels, S. Bradner. The Internet Society, March 1997. ftp://ftp.rfc-editor.org/in-notes/rfc2119.txt
- [VCARD] F. Dawson & T. Howes, vCard MIME Directory Profile, IETF, RFC 2426, September 1998. http://www.ietf.org/rfc/rfc2426.txt
- [DC] The Dublin Core Metadata Initiative http://dublincore.org
- [OMA] Open Mobile Alliance (OMA) Digital Rights Management V2.1 http://www.openmobilealliance.org/Technical/release_program/drm_v2_1.aspx
- [CC] Creative Commons Initiative Licenses http://creativecommons.org/about/licenses/
- [PLUS] Picture Licensing Universal System (PLUS) License Data Format http://www.useplus.com/useplus/license.asp
- [ISO-4217] ISO 4217 currency and funds name and code elements http://www.currency-iso.org/iso_index/iso_tables.htm
- [BCP-47] Tags for Identifying Languages. IETF Best Current Practice, September 2009 http://www.rfc-editor.org/rfc/bcp/bcp47.txt
- [ISO-8601] ISO 8601 – Representation of dates and times http://www.iso.org/iso/date_and_time_format
- [P3P] The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. W3C Recommendation 16 April 2002http://www.w3.org/TR/P3P/
- [W3CXMLSCHEMA] XML Schema Part 2: Datatypes Second Edition. W3C Recommendation 28 October 2004http://www.w3.org/TR/xmlschema-2/
- [ONIX] ONIX for Books – Release 3.0. EDItEUR, April 2009http://www.editeur.org/93/Release-3.0-Downloads/

103 http://www.w3.org/community/odrl/
104 http://www.w3.org/community/odrl/two/vocab/
105 http://creativecommons.org/ns
106 http://dublincore.org/
107 http://www.loc.gov/standards/premis/

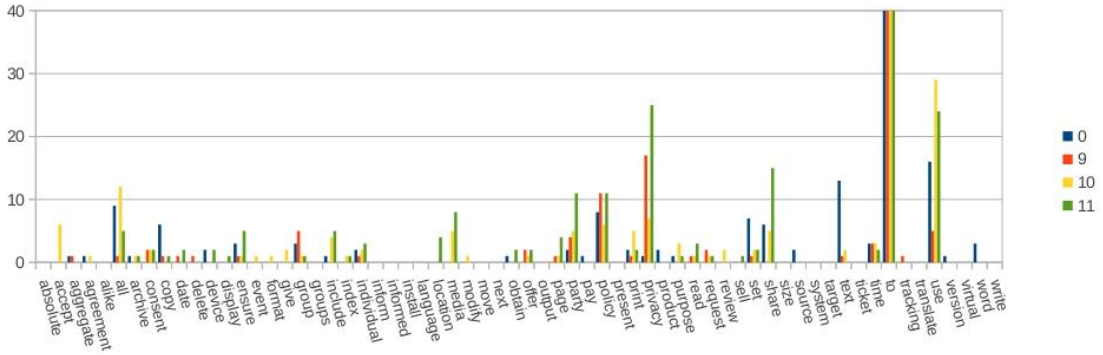


Figure 4.3: Frequency of phrases from the ODRL vocabulary across the Financial Times copyright document (0), privacy policy (9) and Guardian's terms of service (10) and privacy policy (11).

The count of each phrase for the four documents is shown in Figure 4.3. The figure shows, naturally, that the preposition *to* is the most frequent word across all documents, but it also shows the term *use* as a high frequency word in the Guardian's terms of service, indicative of the types of usage allowed in relation to the blog.

4.3.3 Similarity

In this section, we will look at term frequency in relation to the frequency of documents containing the term across a 16 document set including the four documents. The measure known as *tf-idf* is a well-known measure used in information retrieval[108] for assessing the weight of terms with respect to a given document within a corpus. This is likely to show us a better characterisation of the documents that can be compared across the four documents.

In Figure 4.4 we present the tf-idf measure across the general digital rights terms used in Section 2.

108 http://en.wikipedia.org/wiki/Tf%E2%80%93idf

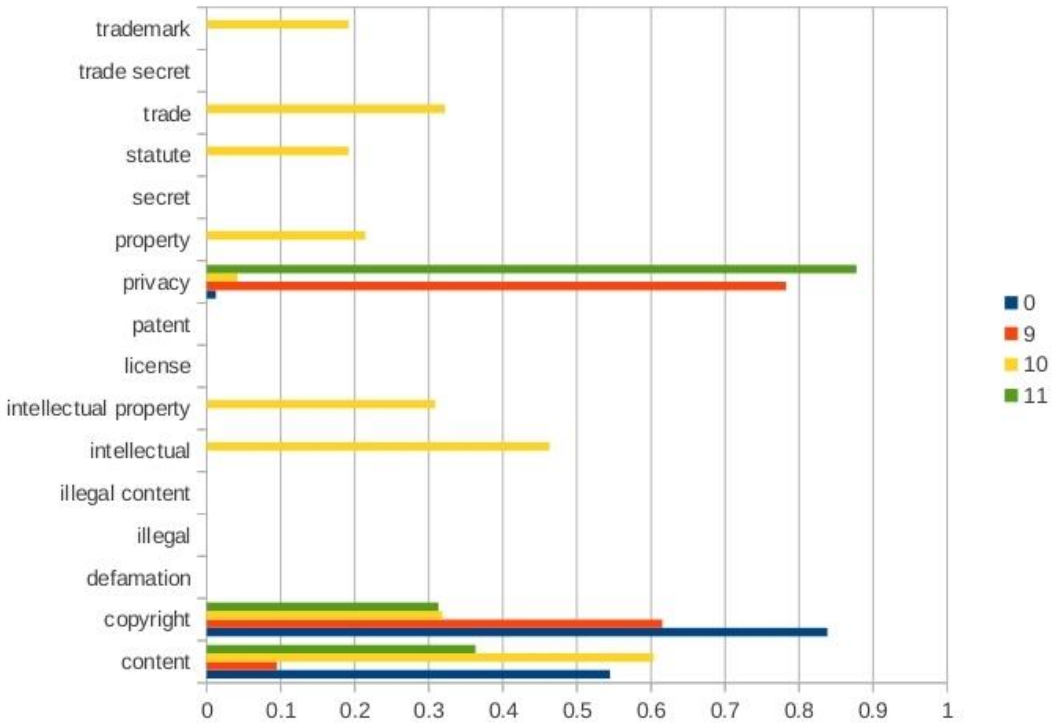


Figure 4.4: Tf-Idf weights of general digital rights terms with respect to the documents 0, 9, 10, and 11.

The results of Figure 4.4 shows that, in contrast to Figure 4.2 where we examined the raw counts of terms *copyright* and *content* are discussed actively in the terms of service and copyright documents, while the term *privacy* is distinctively prominent within the two privacy documents. However, there is also heavy weight on copyright within the Financial Times privacy policy and terms such as *intellectual property* and *trademark* does figure highly in the Guardian terms of service. However, it must be noted that there was no separate copyright document found for Guardian, it may be assumed that they have used the terms of service to specify copyright restrictions rather than the privacy policy document.

In Figure 4.5, we have presented a similar graph with respect to the ODRL vocabulary. This graph seems harder to interpret but, at first glance, it seems that the two privacy documents do tend to talk about consent and/or privacy related terms while the two other documents span across a wide variety of issues concerning one or more of topics surrounding copying, using, sharing and printing.

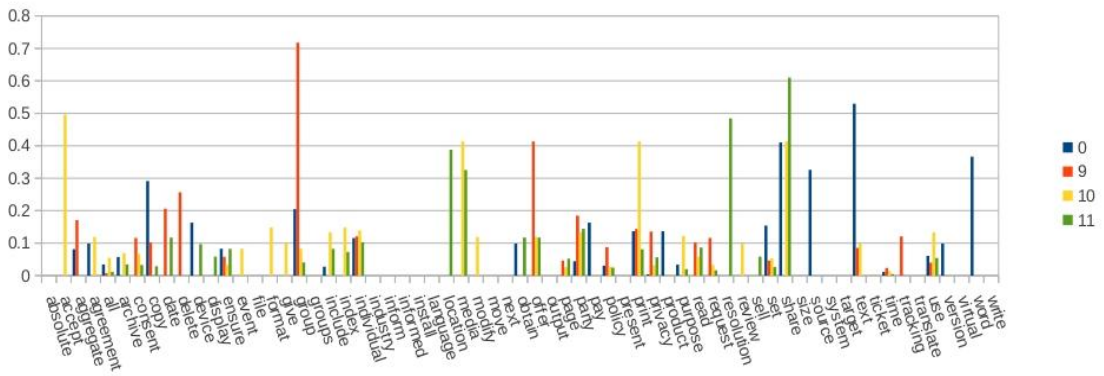


Figure 4.5: Weight of ODRL terms across the documents 0, 9, 10, 11.

The study in this section is too limited to make any firm conclusions or extraction methodology, but the analysis does show that documents do touch on a variety of digital rights management issues raised in Section 2, and further examination might yield more specific locations of target metadata.

4.3.4 Link Analysis

The link analysis in this section is not meant to be comprehensive. The objective of this section is to explore whether the hypertext included in policy data might yield some direction for automated methods of linking the right documents to metadata fields.

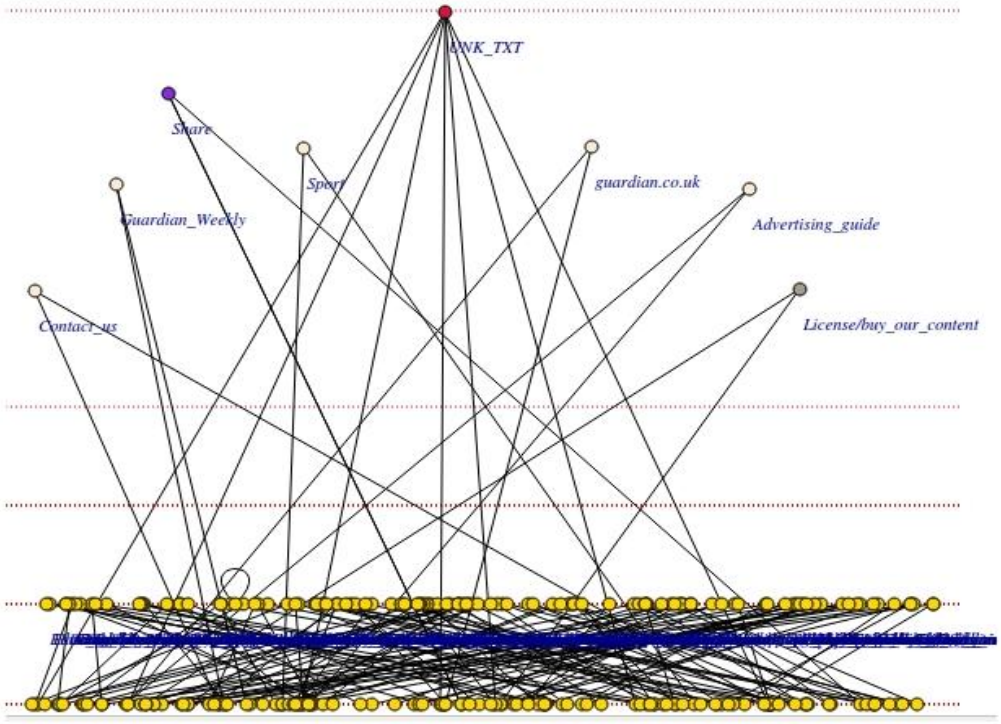


Figure 4.6: Hypertext graph for the Guardian terms of service document. The term "UNK_TXT" means that there was no text extracted by the HTML parser. This may mean that the link was made to an image or a logo.

In Figure 4.6, we present the hypertexts that appear in the Guardian terms of service document. The hypertext is displayed in a graph where each node of the graph represents a unique hypertext. The node, with the most number of URLs linked to it, is displayed at the top as a crimson coloured circle (in this graph the hypertext was set to be "UNK_TXT" because there was no text extracted by the HTML parser). The next popular hypertext (node coloured in violet) is "Share". The figure also shows a node "License/buy our content" suggesting that the content is likely to be regulated by a license.

It seems worth researching patterns of hypertexts further to explore the possibility of using hypertexts to locate rights metadata.

4.4 Conclusions

In this section, we have described the policy data that the BlogForever spider is currently harvesting to support digital rights management. We have also outlined the capability of the repository to

support the actions of curators for collecting and assigning rights metadata. Further, we have explored the possibility of automating metadata generation by looking at the content of the policy data to examine frequent word in relation to ODRL vocabulary and general digital rights terms arising from Section 2.

Even if the documents being retrieved by the BlogForever are adequate for understanding the rights assigned to content, the current process shows that recall of *policy data* is poor: using the current method of retrieval only 1.8% blogs are accompanied by a policy document.

The policy data used in this study is far too small to make any firm conclusions. However, we also observe that:

- blogs managed by the same organisation (e.g. Guardian newspaper) tend to share the same policy documents across all their blogs,
- the retrieved documents do seem to mention 63.72% of the terms of the ODRL vocabulary which promises fertile ground for further exploration,
- the hypertext statistics show that there might be some possibility of taking advantage of these to determine the best links to keep, the best right metadata to use, or to obtain fine grained semantics about rights assigned to content.

There is scope for improving the spider to get better retrieval performance. This is recommended as a future research direction, but it is not clear that this method can be utilised immediately. The final conclusion of this section, however, concurs with the conclusions of other sections in this deliverable: the way forward for collecting, recording, improving, and managing rights metadata lies with pursuing active cooperation between the three main stakeholders of the content (content provider, manager, and user). This involves communicating with content providers to protect their rights and solicit their contribution in shaping rights management (see discussion in Section 2), recognising the rights management opportunities in relation to these at different points of the digital information life-cycle (see recommendations in Section 3), and using best practice metadata standards for assigning rights to content.

In Section 5, we will discuss the recommended metadata schema and cataloguing standards to take this forward.

5 Cataloguing examples for Rights Metadata

This section includes a recap of the rights metadata options available, already stated in D3.1; a note with suggestions on how to implement rights metadata for your instance of BF; a brief description of three types of stakeholders, and their needs; and a description of the three principle types of rights metadata that are relevant (copyright, access, and the right to preserve). The section then provides detailed examples of how these elements could be expressed and catalogued in four standard schemas: MARC XML, Dublin Core, METS Rights, and PREMIS.

5.1 Rights metadata options

Rights metadata are metadata documenting the rights holders, copyright status, permissions, agreements, terms and conditions, and licensing information associated with a blog.
In D3.1 we identified several metadata standards and rights expression languages (RELs) that include fields for statements of digital rights. These allow the expression of rights statements associated with blog content. The standards identified were:

1. CopyrightMD, an XML schema for recording characteristics that, taken together, help determine the copyright status of a resource.
2. METSRights, an extension schema to the METS packaging metadata standard.
3. XrML, a proprietary method for securely specifying and managing rights and conditions.
4. The Open Digital Rights Language (ODRL, 2011) Initiative, an open standard for defining a model and vocabulary for the expression of terms and conditions over assets.
5. Creative Commons Rights Expression Language (ccREL), a proposed Rights Expression Language (REL) for descriptive metadata to be appended to media that is licensed under any of the Creative Commons licenses.
6. Simple Dublin Core, which may be used to describe a resource (DMCI, 2011) and IPR rights attached to one or more digital objects.
7. Qualified Dublin Core, which extends the 15 core descriptive elements of Dublin Core, providing a more granular metadata structure.
8. The PREMIS Data Dictionary (PREMIS, 2011), which includes semantic units for Objects, Events, Agents and Rights.

5.2 Implementing rights metadata

This report believes it is good practice to keep permanent records of rights information associated with the blogs you are harvesting in your organisation. Such records can be referred to in case of any disputes or issues. It is technically possible to express this rights information as metadata. There are various pre-determined metadata schemas available for doing this. It is also possible to store the metadata in a database, thus enabling better management.

In terms of implementing such metadata in BlogForever, "Invenio gives you the freedom to use any tag as long as you document what it means." Our understanding it that this means that custom fields can be built in your instance of the BlogForever database. It also implies that a data dictionary, with documented definitions, must be built.

We recommend using a standard metadata schema to do this. This has the following advantages:

1. The fields have defined names
2. The fields have defined values
3. Your rights metadata will be standardised
4. Your rights metadata will be understood by other repositories, in case of interoperability requirements

5.2.1 Three principal stakeholders

Using rights metadata in a managed way, it will be possible for your instance of BlogForever to meet the needs of three groups of stakeholders:

- Copyright owners (bloggers, contributors, publishers, companies, etc.), who own copyright and intellectual property rights in their blog content.
- Users of content (readers), who have rights of access to the archived blog content.
- Curators / archivists / information professionals, who may be granted the right to store and preserve archived blog content.

From here we can identify three types of rights metadata we need to express.

- Copyright and IPR
- Access
- Right to preserve

5.3 Types of rights metadata

5.3.1 Copyright and IPR

This refers to the ways in which a blog owner declares **copyright**, ownership and intellectual property rights of their content. This will probably take the form of a simple declaration published on the blog itself. It might include the name of a copyright owner, the dates when such copyright applies, the extent of the copyright, and the uses they are prepared to allow of their blog.

If such information is available, you must add it to the record of the blog you're proposing to archive in BlogForever. The curator of your instance of BlogForever should express the copyright information in the database. This information becomes part of your catalogue. You now have a permanent record of the copyright ownership of the blog you have archived.

5.3.2 Access

Access has two dimensions:

- Permissive. The use which an end user (reader / consumer) can lawfully make of the blog content archived in BlogForever.
- Restrictive. The ways in which user access could be restricted. This will often take the form of an automated registry action, such as redaction or use of Technical Protection Measures. It could also reflect a wider policy decision on behalf of your organisation, such as a decision to defer publishing for one year; publish only certain parts of a blog; or restricting access to a specific location (i.e. your reading room).

As to permissive access, there may be a license assigned to you by the original blogger; and the license or licenses which you will publish for your entire blog archive, making declarations that will allow users access to the archived blogs without fear of copyright infringement or breaking the law. If you have prepared a permissions agreement or a consent form (see elsewhere in this deliverable), then you could refer to this agreement in your rights metadata. It's also possible that the blogger has already assigned a license to his blog, perhaps using Creative Commons. It may be possible to reuse or reference that same license in your repository.

As to restricted access, this is best managed by policies, rather than by metadata. Please see the section of this report on best practices.

5.3.3 Right to preserve

Preservation rights are about whether you as archivist / curator have the right to be crawling blog content, storing it on your servers, and preserving it. If you have prepared a permissions agreement or a consent form (see elsewhere), then you could refer to this agreement in your rights metadata.

While Copyright and Access rights metadata ought to be exposed to the public in some way, the preservation rights metadata will probably be for repository use only.

PREMIS is particularly useful for detailed management of preservation rights. The PREMIS schema provides numerous fields to express it and manage it.

5.4 How to express rights metadata within standard schemas

Below we will look at some selected metadata schemas, and propose some very basic catalogue rules that would enable a user of the platform to express the various elements of rights metadata within these schemas. As will become evident, the schemas vary as to the level of detail proposed. There is also some overlap between them; they often identify very similar entities, even when they disagree as to what they are called.

The suggested values column in each table indicates the sort of metadata that might be expressed for rights management purposes. When implementing this, a curator may wish to consider some form of normalisation for the values, or the development of rules that enable greater consistency in the entries.

MARC XML

MARC XML is the de facto schema that Invenio supports. The field names and definitions here come directly from the BlogForever data model (BlogForever: D2.2 Weblog Data Model (2011) [109], p 46). In the Deliverable D4.4, the WP4 team mapped these entities to the MARC XML value 542, "Information Relating to Copyright Status".

Copyright and IPR in MARC XML

Field	Definition	Suggested values
copyright 542	Notes of copyright as retrieved from the blog	© [Full Name] and [Blog Name], [Current Year or Year Range] © [Company Name] and [Blog Name], [Current Year or Year Range]
ownership rights 542	Notes of ownership rights as retrieved from the blog	All rights reserved. Unauthorized use and/or duplication of this material without express and written permission from this blog's author and/or owner is strictly prohibited.

Access in MARC XML

Field	Definition	Suggested values
distribution rights	Notes of distribution rights as	Excerpts and links may be used, provided that

109 http://blogforever.eu/wp-content/uploads/2011/11/BlogForever_D2.2WeblogDataModel.pdf

542	retrieved from the blog	full and clear credit is given to [Name] and [Blog Name] with appropriate and specific direction to the original content.
access rights 542	Notes of access rights as retrieved from the blog	• Permitted • Denied • Content partially open • On-site access only
license 542 $f	Description of an [access] license granted by the copyright owner.	• Standard access license granted by [copyright owner] on [date] via [repository] consent form • Link to a Creative Commons license

Preservation rights in MARC XML

Field	Definition	Suggested values
license 542 $f	Description of a [preservation] license granted by the copyright owner.	Right to preserve granted by [blog owner] on [date] via [name of repository] standard consent form.

Although not detailed within D4.4, there are also subfields available within the 542 entity, some of which may be useful to express further detail of copyright, though not all of them will apply to blogs:

$a - Personal creator
$b - Personal creator death date
$c - Corporate creator
$d - Copyright holder
$e - Copyright holder contact information
$f - Copyright statement
$g - Copyright date
$h - Copyright renewal date
$i - Publication date
$j - Creation date
$k - Publisher
$l - Copyright status
$m - Publication status
$n - Note
$o - Research date
$p - Country of publication or creation
$q - Supplying agency
$r - Jurisdiction of copyright assessment
$s - Source of information
$u - Uniform Resource Identifier
$3 - Materials specified
$6 - Linkage
$8 - Field link and sequence number

5.4.1 Dublin Core

Simple Dublin Core has 15 elements which may be used to describe a resource (DMCI, 2011). IPR and copyright elements can be expressed using four DC fields: Creator, Publisher, Contributor and Rights. Dublin Core has the advantage of being endorsed by the METS Editorial Board.

Using simple Dublin Core, it would be feasible to express all rights management metadata using a single field. Since Dublin Core elements are repeatable, they can be used three times to record different aspects of rights management.

Copyright and IPR in Dublin Core

Field	Definition	Suggested values
<dc:rights>	This field can be used to record information about the date of creation/publication, the owner of the rights, as well as information about the access conditions. Alternatively, the field may contain a URL which points to this.	1. Date(s) of copyright 2. Owner of the rights 3. URL which points to the above

Access in Dublin Core

Field	Definition	Suggested values
<dc:rights>	As above.	• Access conditions • Dates when applicable • URL which points to the above

Preservation rights in Dublin Core

Field	Definition	Suggested values
<dc:rights>	As above.	Preservation rights granted by [blog owner] on [date] via [name of repository] standard consent form.

5.4.2 Qualified Dublin Core

Using qualified Dublin Core, four fields could be used to parse the copyright and licensing information as needed.

Copyright and IPR in Qualified Dublin Core

Field	Definition	Suggested values
<dcterms:rightsHolder>	A person or organisation owning or managing rights over the resource.	© [Full Name] and [Blog Name]
<dcterms:dateCopyrighted>	Date of a statement of copyright.	© [Current Year or Year Range]

Access in Qualified Dublin Core

Field	Definition	Suggested values
<dcterms:accessRights>	Information about who can access the resource or an indication of its security status.	• Permitted • Denied • Content partially open • On-site access only
<dcterms:license>	References a legal document	• Standard access license

	giving official permission to do something with the blog, preferably via a URI. However, this might also be a hard-copy deposit or donation agreement.	granted by [copyright owner] on [date] via [repository] consent form • Link to a Creative Commons license

Preservation rights in Qualified Dublin Core

Field	Definition	Suggested values
<dcterms:license>	References a legal document giving official permission to do something with the blog, preferably via a URI. However, this might also be a hard-copy deposit or donation agreement.	Preservation rights granted by [blog owner] on [date] via [name of repository] standard consent form.

5.4.3 METSRights

Copyright and IPR in METSRights

Field	Definition	Suggested values
<RightsDeclaration>	A broad declaration of the rights associated with a blog or part of a blog intended to inform the user community of these rights.	All rights reserved. Unauthorized use and/or duplication of this material without express and written permission from this blog's author and/or owner is strictly prohibited.
<RightsHolder>	Details of any person or organisation holding some rights to a given blog or part of a blog.	© [Full Name] and [Blog Name] © [Company Name] and [Blog Name]
<Context>	Describes the specific circumstances associated with who has what permissions and constraints.	Excerpts and links may be used, provided that full and clear credit is given to [Name] and [Blog Name] with appropriate and specific direction to the original content.

Access in METSRights

Field	Definition	Suggested values
<RightsDeclaration>	A broad declaration of the rights associated with a blog or part of a blog intended to inform the user community of these rights.	1. Permitted 2. Denied 3. Content partially open 4. On-site access only
<Context>	Describes the specific circumstances associated with who has what permissions and constraints.	Refers to this archived copy, not the original blog

Preservation rights in METSRights

Field	Definition	Suggested values
<RightsDeclaration>	A broad declaration of the rights associated with a blog or part of a blog intended to inform the user community of these rights.	Preservation rights granted by [blog owner] on [date] via [name of repository] standard consent form.

<Context>	Describes the specific circumstances associated with who has what permissions and constraints.	Repository may perform copying, moving, migration, and other actions necessary to preserve the content.

5.4.4 PREMIS

PREMIS is useful for recording the right of the repository to carry out preservation. It's also possible to express copyright and licensing with a very fine-grained degree of accuracy, more so than the other schemas in this section. PREMIS is also unique in allowing you to refer to statutes and legislation that affect rights.

Implementing PREMIS can be complex. It is a holistic method that requires metadata for all the components within a repository (environment, software, people), which means there is an obligation to add links to objects and to agents.

Copyright and IPR in PREMIS

Field	Definition	Suggested values
4.1.2 rightsBasis	Designation of the basis for the right or permission described in the rightsStatementIdentifier.	Copyright
4.1.3 copyrightInformation	Information about the copyright status of the blog.	[N/A: container field]
4.1.3.1 copyrightStatus	Copyright status of the object at the time the rights statement is recorded	copyrighted = Under copyright. publicdomain = In the public domain. unknown = Copyright status of the blog is unknown.
4.1.3.2 copyrightJurisdiction	The country whose copyright laws apply	USA France Germany
4.1.3.3 copyrightStatus DeterminationDate	The date that the copyright status recorded in copyrightStatus was determined	2010 20110908
4.1.3.4 copyrightNote	Additional information	1. Copyright expiration expected in 2015 unless renewed. 2. Copyright statement is embedded in blog footer.
4.1.8 linkingAgentIdentifier	Identification of one or more agents associated with the rights statement.	[N/A: container field]
4.1.8.1 linkingAgentIdentifierType	A designation of the domain in which the linking agent identifier is unique.	An agent in the BlogForever environment.
4.1.8.2 linkingAgentIdentifierValue	The value of the linkingAgentIdentifier.	Name of the copyright owner
4.1.8.3 linkingAgentRole	The role of the agent in relation to the rights statement.	• Blogger • Contributor • Creator • Publisher • Company

Access in PREMIS

Field	Definition	Suggested values
4.1.2 rightsBasis	Designation of the basis for the right or permission described in the rightsStatementIdentifier.	License
4.1.4 licenseInformation	Information about a license or other agreement granting permissions related to a blog.	[N/A: container field]
4.1.4.1 licenseIdentifier	A designation used to identify the granting agreement uniquely within the repository system.	[N/A: container field]
4.1.4.1.1 licenseIdentifierType	A designation of the domain within which the license identifier is unique.	The BlogForever environment
4.1.4.1.2 licenseIdentifierValue	The value of the licenseIdentifier.	UID in BlogForever
4.1.4.2 licenseTerms	Text describing the license or agreement by which permission was granted.	Reference to your standard consent form or standard license.
4.1.4.3 licenseNote	Additional information about the license.	Other types of information related to the license, such as contact persons, action dates, or interpretations.
4.1.8 linkingAgentIdentifier	Identification of one or more agents associated with the rights statement.	[N/A: container field]
4.1.8.1 linkingAgentIdentifierType	A designation of the domain in which the linking agent identifier is unique.	An agent in the BlogForever environment.
4.1.8.2 linkingAgentIdentifierValue	The value of the linkingAgentIdentifier.	Name of the grantor of the license
4.1.8.3 linkingAgentRole	The role of the agent in relation to the rights statement.	Grantor

Preservation rights in PREMIS

Field	Definition	Suggested values
4.1 rightsStatement	Documentation of the repository's right to perform one or more acts.	[N/A: container field]
4.1.1 rightsStatement Identifier	The designation used to uniquely identify the rights statement within a preservation repository system.	[N/A: container field]
4.1.1.1 rightsStatementIdentifier Type	A designation of the domain within which the rights statement identifier is unique.	The BlogForever environment
4.1.1.2 rightsStatementIdentifier Value	The value of the rightsStatementIdentifier.	UID
4.1.6 rightsGranted	The action(s) that the blog owner has allowed the repository.	[N/A: container field]
4.1.6.1 act	Preservation actions the repository can take	• Copy • Migrate • Modify • Use

		• Disseminate • Delete
4.1.6.2 restriction	Limitations on the acts	• No more than three copies • Allowed only after one year of archival retention • Blogger must be notified
4.1.6.3 termOfGrant	Time period for the permissions granted.	[N/A: container field]
4.1.6.3.1 startDate	Start date for permission granted.	20110909
4.1.6.3.2 endDate	End date for permission granted.	20150807
4.1.6.4 rightsGrantedNote	Additional information about rights granted.	This semantic unit may include a statement about risk assessment, for example, when a repository is not certain about what permissions have been granted.
4.1.7 linkingObjectIdentifier	The identifier of a blog associated with the rights statement.	[N/A: container field]
4.1.7.1 linkingObjectIdentifierType	A designation of the domain in which the linking object identifier is unique.	The BlogForever environment.
4.1.7.2 linkingObjectIdentifierValue	The value of the linkingObjectIdentifier.	Value of the blog object (its UID in BlogForever).
4.1.8 linkingAgentIdentifier	Identification of one or more agents associated with the rights statement.	[N/A: container field]
4.1.8.1 linkingAgentIdentifierType	A designation of the domain in which the linking agent identifier is unique.	An agent in the BlogForever environment.
4.1.8.2 linkingAgentIdentifierValue	The value of the linkingAgentIdentifier.	Name of the archivist or curator.
4.1.8.3 linkingAgentRole	The role of the agent in relation to the rights statement.	Archivist.

6 Interviews

As a complementary approach to the above mentioned strategies for developing rights management policies in the archiving of weblogs, the BlogForever project partners involved in Task 3.3 conducted a series of interviews with experts involved in the field of digital archiving or the collection of digital material on the web. These individuals were identified as being knowledgeable in one of the several key areas listed in chapters 2 and 3 of this document, and thus able to provide valuable insight for our work. Section 6.1 describes the methodologies used to identify interview subjects, gather data through the use of qualitative interviews, analyse the data and sort it into relevant categories. Section 6.2 presents the highlights of the interviews conducted and some of the discussions that emerged as a result of those interviews. The final section, 6.3, summarises the conclusions we have drawn from the interviews conducted and what those conclusions mean for the objectives of the Task 3.3 as described in the project's DoW.

6.1 Methodology

Expert interviews are typically applied as instruments in qualitative research (Flick 2009):

- to provide comparison or express variance of opinion/knowledge in a given field
- to gain orientation in a subject for the development of further research instruments (such as questionnaires or task-lists for focus groups)
- to compliment, or "round out" other interviews
- to validate findings

As such, expert interviews are rarely "stand-alone" methods of data collection, but rather part of a larger strategy in data collection, analysis and validation.

The decision to use qualitative interviews to supplement the research undertaken as part of D3.3 was made with two objectives in mind: The first was to illustrate the complexity of rights management issues pertinent to the archiving of weblogs. The second was to validate our findings, in particular the recommended strategies and treatments for managing rights issues within the context of archiving weblogs. It is worthwhile to note that these interviews were conducted to support our research and do not represent the main body of this deliverable. The following sub-sections will assist the reader in understanding how information from these individuals was collected and how it was analysed and used during the development of this document.

In each section, we also refer to the relevant limitations of this method: difficulty in identifying the right experts, time and scheduling restrictions, demand of high level of expertise and confidentiality (Flick 2009).

6.1.1 The Sample

The sample of interview partners chosen for this task can be described as non-probabilistic, based on purposeful and convenience sampling techniques.

A non-probabilistic sample is generally more appropriate for qualitative research, as random samples produce significant sampling errors in studies of small scope (Marshall 1995). For our purposes, a small sample of interview partners was sufficient to meet our objectives as described above. As such, we intended to interview between 3 and 6 individuals, depending on the range of responses. As sociologist Howard Becker noted in "How many qualitative interviews is enough?", a paper published by the National Centre of Research Methods in the UK, even one interview can suffice if the goal of the interview process is to illuminate complexity of an issue (Baker & Edwards

2012). We found that our needs were met after conducting 3 interviews, the highlights of which are presented below.

The purposeful and convenience sampling techniques we applied are visible in how we specifically identified and secured interview subjects. All potential interview subjects were identified first by means of their connections to certain general topics that are related to rights management in the context of web archiving. These topics include: copyright, intellectual property rights, defamation and liability, crawling and other types of ingest strategies and privacy. In addition, we sought out individuals that had a broad range of expertise specifically in archiving science and the various collection management activities that belong to this field. After conducting two initial interviews of a more general nature, we reviewed the transcripts and identified some more detailed areas of attention in which it would be necessary to conduct a second round of interviews. We consulted our professional networks to help us engage with individuals who fit our sampling frame and interviewed those who agreed to be interviewed.

These sampling techniques have both benefits and complications. The main benefit of purposeful sampling was that it provided us with enough insight to shape our research and draw some important conclusions early on (see Discussion and Highlights), saving the project a considerable amount of time and effort. The complications of the method were limited to interview subject availability and desire for anonymity. Concerning the first issue, it is time consuming to arrange an appropriate interview setting and time with subjects who are experts in their field. Their schedules were less flexible and the interviewer was required to be very spontaneous. The use of convenience sampling was a response to this limitation. The second issue is more complex and is visible in the sub-section "Discussion and Highlights" below. When an interview subject was connected with a particular institution, they were cautious in speaking about any legal issues that could potentially put their institution at risk. All interview subjects were offered anonymity and, for this reason, we are unable to reveal the full details of their expert knowledge. Additionally, any revealing details about the research subjects are omitted from the transcripts, making it difficult to perform a secondary analysis of the data collected by external parties (Wiles 2012).

The final 3 interview subjects can be described in the following way[110]:

1. "Paul Johnson" - a leading consultant and expert in the field of Information Technology and Law, working with major European digital collections on rights management issues.
2. "Gary Fields" - a Web Archiving Engagement and Liaison Officer at a large and influential institution maintaining a collection of digital works.
3. "Sam Howard" - a business executive in the field of content ownership, with 25 years of experience in content licensing of news (both paper and web)

6.1.2 Research Conditions

All interviews were conducted either over telephone or using voice-over-IP technologies (in particular, Skype). In most cases, the interviews were recorded, only after confirming consent from the interview subject. If the subject did not give his or her consent, the interviewer made notes during the interview to whatever extent possible and created a summary of the interview. This summary was then sent to the interview subject for confirmation and/or clarification before being absorbed into the data pool. The interviewers were chosen from the BlogForever project partners involved in D3.3.

These research conditions presented some advantages and challenges relative to validity. On the whole, the use of telecommunications for interviewing experts was very advantageous. This technique allowed us to be more flexible with scheduling and connect with individuals from outside

110 All names have been changed and identifying information removed

of the geographic reach of the BlogForever project partners. In addition, the use of Skype made it possible to build the kind of rapport with interview subjects that is typically associated with face-to-face interviews (Hanna 2012). Recording the interviews assisted with the analysis of the data, in particular with the thematic coding that lead us to identify the critical categories for the second round of interviews. The issue of having different interviewers, with differing levels of expertise in web archiving and associated legal concerns, was a challenge in that the interviewer did not always fully understand the answer of an interview subject enough to ask relevant follow-up questions. In all cases, we were able to contact the interview subject at a later date for clarification, mitigating potential validity risks. On one occasion, the interviewer conducted a more informal interview, during which the interview itself was not recorded and the questions asked were spontaneous (see subsection 8.1.3 "Research Instruments" below). While this may have affected the continuity of the interviews, it was a necessary method for producing a positive rapport with that specific interview subject and working with his availability to be interviewed. As we did not aim to reach saturation or complete coverage, we do not feel that this influenced the overall value of conducting the interviews..

6.1.3 Research Instruments

The majority of interviews were conducted with the guide of interview questions, which can be viewed in Appendix C of this document. This type of research instrument allows the interviewer to conduct a semi-structured interview during which he or she is able "to collect qualitative data by setting up a situation (the interview) that allows a respondent the time and scope to talk about their opinions on a particular subject."

This method was chosen for its flexibility in discussing dynamic topics with high complexity. In entering the interview with less pre-judgement on behalf of the interviewer about which aspects of the topic were of greatest importance, the interview subjects were more freely able to express their own opinions, thereby increasing the validity of data collected. The instrument also allowed interviewers with less interview experience or technological knowledge to discuss topics with the interview subject more comfortably.

One potential challenge of using semi-structured interview guides was that all interviews were essentially different and not all of the interview subjects were given the same questions to answer. While this would technically affect the standardisation of such interviews, our goal was to tailor interviews to the specific field of expertise of each interview subject.

6.1.4 Sorting and Analysis

Analysis of the interviews was conducted with the use of generalised and thematic coding, on interview transcripts (which can be viewed in Appendix E). Thematic coding involves the study of transcripts for particular patterns in responses, as well the mention of specific subjects deemed relevant to a given research question.[111] The main benefit of such an approach was that BlogForever partners with more technical or legal knowledge were able to review the transcripts and assist the interviewers in extracting the most relevant information from the data. The challenges of this approach are those that are generally associated with qualitative data collection and analysis, in particular that the analysis can be viewed as subjective and "lacking in precision".[112] However, the analysis was sufficient to meet the goals of this deliverable, in that it provided BlogForever partners involved in task 3.3 with adequate data to support our claim that rights management is an extremely complex area of research and that our proposed treatments for addressing rights management in the archiving of weblogs address the main concerns in a realistic and practical way.

111 http://isites.harvard.edu/icb/icb.do?keyword=qualitative&pageid=icb.page340897

112 http://www.edu.plymouth.ac.uk/resined/qualitative%20methods%202/qualrshm.htm#Qualitative Research Assessed

6.2 Discussions and Highlights

Within each semi-structured interview conducted there were two main threads of discussion: The first involved the expert's professional insight into known issues concerning rights management in the context of digital preservation. The second related to the expert's own sense of the most critical challenges and opportunities facing the digital archiving community (and related communities of digital content aggregators) relative to rights management. The results of these conversations are presented below arranged by theme for the purposes of coherence and readability. For the original transcripts please see Appendix B.

6.2.1 Legal Risks Associated with Rights Management in Digital Archiving

All respondents agreed that copyright, intellectual property rights, privacy and data protection, and defamation/libel are the key rights issues for digital archives, along with the potential for archiving material that is illegal in and of itself (such as child pornography). The main issue for all respondents was assessing the *level* of risk involved in certain activities that belong to digital preservation (or aggregation). For respondents, assessing risk related to a number of factors.

First, one must consider the specific political and legal environments in which digital preservation or aggregation occur. One respondent, PW, mentioned the Danish National Library and the importance of data protection under Danish law. According to the respondent, data protection is a more sensitive issue in Denmark, which may explain why the Danish National Library requires users to engage with the digital collections "on site". The same respondent spoke about privacy and data protection in Sweden, as evidenced by the Lindqvist case.[113] In this case, the publication of sensitive data about co-workers on the personal website of the defendant was deemed as publication and in violation of the privacy and data protection rights of the individuals whose personal details she made public. To the respondents' knowledge, no similar case has yet been brought to court in the UK and was also not likely to occur, due to the political and legal environment of the country. However, two respondents did mention a notable case with regard to defamation law in the UK: (here is where I would explain how it works in UK and what happened in the law suit). In this case, it was ruled that every time someone accessed the webpage, it counted as a new publication and thusly a new count of defamation. The Defamation Bill (passed into law as the Defamation Act in April 2013[114]) in the UK was proposed to change this. Under the Defamation Act, there is a now a "single publication rule", which means that as long as the item has not changed materially, any subsequent publications of the material are deemed to be part of the original single publication. The Defamation Act has given the archiving and preservation community in the UK some additional "defenses" against claims of significant damages. Likewise, in consideration of "orphan works", works for which no owner can be immediately identified, the Hargreaves review[115]

Additionally, multi-jurisdiction issues, in cases in which an infringement claim is made across geo-political borders, would also influence the extent to which a particular law could and would be likely to be enforced. Respondents seemed to feel that for copyright, this would be relatively clear, based upon the fact that copyright is general recognized across borders. If, for example, an individual from France owns content that is then published in another country without their permission, that individual should technically be able to seek legal redress if the country where the material was published recognises the rights of the content owner. The question of whether or not legal action is actually brought against the infringing party, depends on the character of the content

113 http://curia.europa.eu/juris/document/document.jsf?text=&docid=48382&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=449594

114 http://www.legislation.gov.uk/ukpga/2013/26/contents/enacted

115 http://en.wikipedia.org/wiki/Hargreaves_Review_of_Intellectual_Property_and_Growth#Plans_for_the_implementation_of_recommendations

owner. The example given by one respondent was that of George Lucas Films. In the case where any material owned by the company is published anywhere without its permission, the company can and has been known to seek legal retribution. However, the issue of defamation/libel was less clear to one respondent.

The character of the institution itself was also listed as a major factor in addressing the level of risk involved in digital preservation, with regards to rights management. Memory institutions with a reputation to uphold as a national body, will tend to be more risk averse than those like the Internet Archive, which have no particular government backing. As such, their policies around seeking permissions, ingesting content and removing infringing materials will be more straightforward.

Of course, some risks are difficult to identify, simply because the law has not caught up with the advancement of technology and the ways in which content is produced and published. Laws that refer to paper works are difficult to translate in a digital context. Without the law explicitly allowing the treatment of digital material in certain ways, there are few courses of action left but to either proceed carefully, or hold materials in a "dark archive" (unpublished archive) until the law changes or copyright expires. One respondent discussed the BBC's "Doomsday Project" in this respect. Due to the fact that more than 1 million people made contributions to the project, the copyright issues are so significant that a large portion of the project will remain in a dark archive until 2090.[116] Laws are also constantly changing. One respondent did not feel that he could even adequately describe what the major challenges are to digital preservation, simply because the laws were continuously both addressing and creating new problems. Additionally, there is the issue of what constitutes "publishing" in the context of digital material. One of our respondents noted the Meltwater case in the US[117], which addressed the issue of crawling and lawfully browsing material. According to this respondent, the acts of crawling and caching of potentially copyrighted material could be considered illegal in the strictness sense, but that the law has had to be interpreted more broadly and pragmatically in this regard. Laws regarding fair use of materials are consistently evolving. The respondent provided the example of fair use toward news items and how the length of the abstract and headlines are important in defining what this means.

Of course, one of the main factors in assessing risk is quantifying it, which is discussed in the section below.

6.2.2 Quantification of Risk

The translation of legal risks associated with rights management into real numbers is relevant for any institution engaging in digital archiving. How often are institutions archiving digital material actually sued or threatened with a lawsuit by a content owner? What is the extent of damages? Our intention in asking this question was to find out what the *perception* of real risk might be for our experts. Two of our respondents working with memory institutions felt that this issue was difficult to quantify, mostly because there have not been enough cases to warrant the collection of such data. However, both of these respondents also agreed that the infrequency of legal claims was also related to the relatively conservative ingest strategies of the institutions with which they are involved. If one applies a permissions-based ingest strategy, for example, in which the content owner is explicitly asked for consent to archive their material, complaints should be minimal. When asked under what situations a complaint would arise (in consideration of a very conservative ingest strategy), the same two respondents agreed that the most likely scenario would be a content depositor, not aware of his or her full rights to the material they deposited, who infringes upon the copyright of a third party. In this case, the third party who owned part of the material archived, would register a complaint. Both respondents also agreed that in most cases, unless the content owner is a professional artist or publishing house, in is unlikely that a complaint would respond in

116 http://en.wikipedia.org/wiki/BBC_Domesday_Project
117 http://www.mpa.org.uk/news/meltwater-case-muddies-the-water-on-content-copyright-online

legal action. Rather, a notice of take-down would be filed, the archive would comply and the issue would be resolved.

One possible way of testing this theory, would be to identify the number of legal actions brought against an institution such as the Internet Archive, which does not seek permission of content owners before archiving material. Rather, the Internet Archive operates under an "opt-out" policy for the ingest of materials. If an individual does not want his or her content to be preserved in the archive, a notice of take down can be registered with the archive and the material is removed.[118] Alternatively, an individual can utilize a piece of code known as "robots.txt" exclusion protocol to communicate to web crawlers that the material should not be visited by the crawler.[119][120]At the time of writing, we were unable to contact the Internet Archive for a response, but the general feeling among the respondents was that this number would be relatively small. Secondary research would indicate that the Internet Archive has had only a few legal battles, none of which resulted in significant claims.[121]

An opt-out policy does carry with it more potential legal risk. According to two respondents, the Internet Archive could be perceived as being entirely in violation of copyright solely on that basis. In referring to the robots.txt exclusion, one respondent noted that, while robots.txt is a good ad hoc solution to prevent archiving material that an owner does not wish to archive, it does not "include the concept of licensing." However, the same respondent said that, while he felt that explicit permissions should be sought to whatever extent is possible (especially for commercial use), it would not be practical to seek agreement in every single case within the context of web crawling. According to this respondent, under a strict definition of rights infringement, the entire concept of a world wide web (which "copies" and "republishes" under the most technical definitions as its major function) would be illegal. Moreover, in the case of blogs, the respondent felt that connecting to a ping-server, for example, was an explicit agreement on behalf of the content owner (the blogger) to allow his or her material at least to be crawled, if not preserved. Still, there are other risks of an opt-out ingest policy that do not have to do with copyrighted materials at all, regarding the ingest of materials which are in any case illegal to publish, such as some types of pornography. One respondent described how legal actions in the case of owning or distributing such materials are often based around the defendant's knowledge of having done so. For example, if an institution requires permission from content owners to preserve content, but inadvertently preserves illegal content, that institution can refer to their terms and conditions or depositor agreement to illustrate the responsibility of the depositor to ensure that no content provided is illegal. The Internet Archive handles this issue with the following statement:

> "the Collections may contain information that might be deemed offensive, disturbing, pornographic, racist, sexist, bizarre, misleading, fraudulent, or otherwise objectionable. The Archive does not endorse or sponsor any content in the Collections, nor does it guarantee or warrant that the content available in the Collections is accurate, complete, non-infringing, or legally accessible in your jurisdiction, and you agree that you are solely responsible for abiding by all laws and regulations that may be applicable to the viewing of the content. "[122]

Again, it is difficult to know the extent to which such a statement would hold up in a court of law. At the time of writing, no legal claims against the Internet Archive to this effect were identified.

In the experience of respondents who worked directly with memory institutions, content owners are mostly concerned with having their complaints heard and efficiently addressed. As part of the legal claim of damages regarding digital rights, a court will consider the extent to which the infringing

118 http://archive.org/about/terms.php
119 http://www.robotstxt.org/robotstxt.html
120 http://archive.org/about/exclude.php
121 http://en.wikipedia.org/wiki/Internet_Archive#Controversies_and_legal_disputes
122 http://archive.org/about/terms.php

party attempted to deal effectively with a complaint. Metrics such as how long an institution took to respond to an issue, how long information was made available to the public and how many downloads of the information occurred over a given time are important in proving the extent of damages that can be realistically claimed.

One of the ways in which respondents felt that risks could be mitigated was by having clear, firm agreements with content owners/depositors that outline exactly what is allowed to be preserved, what can be done with the material in order to preserve it and how that information can be used (for research or commercial purposes, for example). In the case of one respondent, he felt that even when content would be changed significantly in terms of look and feel for the purpose of preservation (as is the case in the BlogForever repository software component), this could be easily clarified with a depositor as one of the terms of agreement.

6.2.3 Public Perception and Digital Rights

The issue of public perception of digital rights was brought up in relation to two main issues: firstly, that the public has a general, though limited, concept of digital rights. Secondly, that the public can and should be engaged in these issues when considering the preservation of digital content such as weblogs or other personal websites (also micro-blogs like Twitter or Facebook). In consideration of the first issue, one respondent felt that individuals were gaining awareness for how their material can be used and is used by various communities (such as archivists, researchers, companies, etc.) and they do want to be credited for their content. In addition, even when a content owner publishes something widely visible on the internet, this does not necessarily mean that they would like it to be preserved over the long term or agree for it to be used in any other way. One example that he provided involved the use of photographs from Instagram for commercial use.[123]After the community of users placed pressure on Instagram regarding some of the language set forth in their terms of service, Instagram responded with a statement clarifying how and when users' photos could be displayed, or re-produced by Instagram or other third parties.[124]In this example, only a handful of users threatened to take legal action and most of these were professional photographers.

However, the same respondent felt that even if the general public was not likely to be able to follow up on claims of infringement (for financial reasons), a positive relationship with the general public could still help ease the process of preservation and make it easier for archivists to do their jobs in a meaningful way. In his view, the general public still lacks knowledge about what can be done with specific content once it has already been published to the web and what general purpose preservation of such material serves. Thus, he felt that a large part of any digital rights strategy for (at least) memory institutions would have to involve educating the public on certain rights related and preservation issues. One of the strategies he recommended for involving the public was collaboration with service providers such as Facebook or Instagram to both educate the public about the importance of web heritage and gain their support in preparing digital content in such a way as facilitates preservation. Asking content owners to "tag" their work would be one example for how the public can be engaged in producing metadata for content preservation, reducing the amount of effort a memory institution would have to invest in this task. This also becomes more important when archiving personal websites, blogs and other digital information that could contain the work of many different people. Engaging the public in helping to identify content ownership (as well as assigning meta-data) using the channels with which they are already familiar (such as their blog service provider, Facebook or other entity), is the most sensible approach. It also adds a layer of protection to memory institutions from potential legal risks by placing the responsibility with content owners or depositors to know which content is theirs and express how it should be used.

123 http://news.msn.com/world/uks-instagram-act-will-your-photos-become-everyone-elses
124 http://www.theverge.com/2012/12/18/3780158/instagrams-new-terms-of-service-what-they-really-mean

Moreover, as one respondent noted, the public can and should be educated about the service that preservation provides, in terms of archiving valuable information about our understanding of history and social media. As an example of this, he cited the events in Turkey that took place in May of 2013[125], in which social media played an enormous role.

6.2.4 The Future of Digital Rights Management in Digital Preservation

In asking interview participants to predict the direction of digital rights management in digital preservation, we intended to discover the points of development and innovation that the respondent perceived. For all respondents, *gaining public support for preservation should be a top priority*. Not only would this provide archiving institutions with the chance to benefit from the public's contribution to the collection of data, but also to hear from the public what they believe is worthy of preservation. One respondent described how we are quickly approaching the possibility of "saving everything" and that it would become more and more critical to consider what it is that we actually do want to preserve and why. However, he also believed that this role would continue to be played primarily by those who currently do so, archivists, curators, librarians, etc. In terms of the expectation regarding the legal aspects of preservation, all respondents also agreed that it was *unlikely that pure clarification on these points would be achieved*. One respondent noted that the lack of standardization, regarding international law, would continue to contribute to the inability of archiving institutions to achieve risk free strategies around rights management. While some legal battles are fought on national fronts to improve flexibility in areas such as copyright infringement, data protection, defamation and fair use, *this will initially open up additional questions and public discourse*.

6.3 Conclusions

In consideration of the vast and complicated terrain that is rights management, in which national and international laws play a considerable role in determining what actions can be performed legally on digital content, our respondents agreed with our own findings that the task of any institution participating in preservation is firstly, to assess risk. From our respondents, we know that there are *legal risks*, *economic risks*, *social risks* and *preservation risks* involved in this decision making process. If an institution wishes to comply with the legal standards in their country for managing rights issues within their practice (and we recommend that they do so), they will need to *inform themselves about the current legal issues surrounding copyright, defamation, intellectual property, fair use, privacy and data protection*.

Rights management is present within all preservation activities and we have provided some examples of *where to locate risks and manage them* in the previous sections of this document. Even if an institution is able to limit legal risks, these strategies are time consuming and there will still be cases in which mitigation strategies will not be sufficient. Therefore, an institution involved in preservation will need to *consider the resources required to comply with legal standards and follow up on take-down procedures or other types of complaints, formal and informal*. To assist with this practice, a strategy for:

- engaging the public in understanding their rights (and responsibilities), as well as,
- information on how they can contribute to the preservation of digital material,

will be a necessary component of any approach to rights management. Finally, an institution should develop an understanding of how all of the above impacts the integrity of collections and consider possibilities for dealing with such consequences.

125 http://en.wikipedia.org/wiki/2013_protests_in_Turkey

7 Conclusions

In this report, we have presented information to support the development of a digital rights policy that is expected to be a key component in future blog repositories using the BlogForever platform. We have given an overview of core areas of concern (Section 2) which has uncovered the grey areas of intellectual property rights in the web environment, the risks involving invasion of privacy, defamation, and illegal activity and content, and, the relationship between these issues and licenses that can be drafted to explicitly define permitted content and related use. The topics covered in Section 2 are summarised again in a mind map in Figure 7.1.

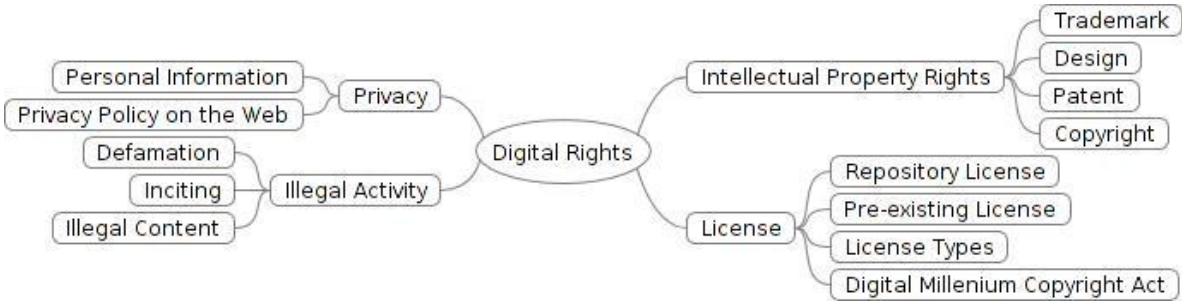


Figure 7.1: Dimensions of digital rights to be considered in the development of a repository rights management policy.

The section ends with a set of conclusions about particular actions that will have to be taken or addressed in light of the risks that have emerged from the discussion (Section 2.6) .The discussions in each of the subsequent sections address these actions by presenting suggested actions in light the observed risks.

For example, we have highlighted explicit points within the digital information life-cycle at which action can be taken to prescribe digital rights management activities, indicating anticipated outcomes (Section 3). This is examined within the context of existing archives to suggest best recommended practices as a starting framework for developing a digital rights management policy (Section 3.5) to be tailored in a way that is most appropriate in light of repository objectives and organisational policies. In fact, if the repository sits within a bigger organisation, a rights management policy may already exist and should be consulted.

In Section 4, we have explored functionalities that are available as part of the BlogForever spider and repository that might serve to help us express the rights management policy. We examine the types of data source that the spider can be configured to retrieve in support of extracting rights metadata and examine the basic rights metadata fields that are derived from the BlogForever data model (derived from the project deliverable BlogForever: D2.2 Weblog Data Model (2011)) to ensure minimum rights management functionality. We also discuss the extensibility of BlogForever for rights management by answering a series of frequently asked questions likely to arise from future repository managers and digital curators of blogs. This is intended to complement the examples we provide in Section 5 to demonstrate how to catalogue rights metadata.

To scope the changing discussions in digital rights management, in Section 6, we have examined themed interviews with experts in the field. These types of interviews are expected to help future curators identify gaps in the general understanding of current discussions in digital rights management. It is recommended that such interviews be employed as part of the risk management activity of a repository in conjunction with internal risk management processes.

The investigation presented in this report leads us to conclude that a digital rights management policy should be based on four core dimensions: **good communication** with the stakeholders (right holders, content users, information professionals), an **open policy**, that is, being explicit about the

responsibility of each stakeholder and how the rights and responsibilities of each stakeholder will be collected and expressed, a **succession plan** (plans regarding what will happen to the content and associated data if the repository cannot continue), and keeping updated about **changes that affect right management** (law, climate, policies, technology, terms & conditions). These four areas are summarised in Figure 7.2 and elaborated according to topics raised with each section of this report.

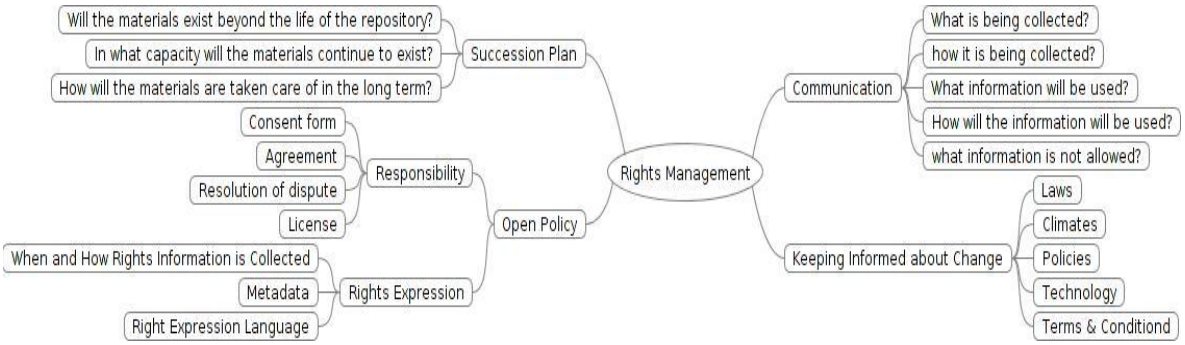


Figure 7.2: Four areas to be addressed in digital rights management.

To conclude this report, we return to question 4 of Section 1.1:

> "What approaches for rights management might be developed in the future?"

This is a difficult question especially in light of the fact that legislation is in flux in response to the social media revolution and the way we use the Internet. However, the conclusion of Section 6 captures some salient points in relation to this issue. The overall opinion that comes across from the interviews of Section 6 is that digital rights management is a process of assessing and managing risks. These include legal, economic, social, and preservation risks. For the curators of information to keep in step with the changing times and legal climate, they will need to engage in a repeating cycle of informing themselves about the current legal issues surrounding copyright, defamation, intellectual property, fair use, privacy and data protection. Perhaps, more importantly, however, future digital rights management may need to involve an effective strategy for:

- engaging the public in understanding their rights (and responsibilities), as well as,
- building information on how they can contribute to the preservation of digital material.

The key is in communication and involvement. The public will be mobilised if they see the benefit of preservation, curation, and innovation as long as they are not kept out of the picture.

In a world where crowd sourcing is suggested as a way of benefiting research through the contribution of the general public, where citizen science allows a lay person to engage in serious scientific research, and, where crowd funding allows the communal funding of innovative projects, *social curation* seems like a natural concept emerging. In fact, initiatives such as Pinterest[126] can be said to be *personal curation in action.* The question, therefore, might lie in communally driven rights management.

126 https://pinterest.com/

8 References

Baker, S. E. and Edwards, R. (2012) "How many qualitative interviews is enough?", National Centre for Research Methods Review Paper, National Centre for Research Methods, 168. http://eprints.ncrm.ac.uk/2273/4/how_many_interviews.pdf

Charlesworth, A. (2009) "Digital Lives >> Legal & Ethical Issues." Digital Lives Research Paper, British Library, 14 October 2009. http://britishlibrary.typepad.co.uk/files/digital-lives-legal-ethical.pdf

Collier, Piccariello & Robson (2004) "A digital rights management ecosystem model for the education community." Report from the Eduworks Corporation. http://www.researchgate.net/publication/228804691_A_digital_rights_management_ecosystem_model_for_the_education_community

Coyle, K. (2006) "Rights in the PREMIS Data Model." A report for the Library of Congress. http://www.loc.gov/standards/premis/Rights-in-the-PREMIS-Data-Model.pdf

CCSDS (2002) *Reference Model for an Open Archival Information System (OAIS)*, Consultative Committee for Space Data Systems, 2002.

DELOS (2007) *The DELOS Digital Library Reference Model.* Report from the DELOS Network of Excellence. http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_096.pdf

Deromedi, N. and Shallcross, M. (2011) *The University of Michigan Web Archives: Collection Development Policy and Methodology Version 1.1.* http://bentley.umich.edu/uarphome/webarchives/UM_WebArchives_Policy_20110324.pdf

DL.org (2010) *The Digital Library Reference Model.* Report from DL.org Coordination Action. http://www.dlorg.eu/uploads/DL%20Reference%20Models/The%20Digital%20Library%20Reference%20Model_v1.0.pdf

Flick, U. (2009) *An Introduction to Qualitative Research.* Fourth Edition. London/Thousand Oaks, California/Dehli. Sage Publications, 168

Hanna, P. (2012) "Using internet technologies (such as Skype) as a research medium: a research note", Qualitative Research, 12 (2), ISSN 1468-7941.

Henderson, S.C. and Snyder, C.A. (1999) "Personal information privacy: implications for MIS managers." *Information & Management* 36 (1999) 213-220. https://www.comp.glam.ac.uk/blackboardAT/IS/IS3S04/CourseMaterial/personal%20data%20implications.pdf

Kim et al, *BlogForever*: *D3.1 Preservation Strategy* (30/09/2012)

Llopis / Encinar, *BlogForever: D4.8 Final BlogForever Platform* (August 2013)

Korn, N. and Oppenheim, C. (2006). "Creative Commons licences in higher and further education: Do we care?" *Ariadne*, 49 http://www.ariadne.ac.uk/issue49/korn-oppenheim/

Koerbin, P (2004) "Managing Web Archiving in Australia: A Case Study." *4th International Web Archiving Workshop*

Marshall, M. N. (1995) "Sampling for Qualitative Research.", *Family Practice*, Vol. 13, 522-525. http://blsciblogs.baruch.cuny.edu/com9640/files/2010/08/qualsampling.pdf

ODRL-MODEL (2012) *Open Digital Rights Language (ODRL) Version 2.0 – Core Model.* Final Specification, W3C ODRL Community Group, 19 April 2012. Iannella, R., Guth, S., Paehler, D. and Kasten, A (eds). http://www.w3.org/community/odrl/two/model/

ODRL-11 (2002) Open Digital Rights Language (ODRL) Version 1.1. W3C Note, 19 Sept 2002. Iannella, R. (ed). http://www.w3.org/TR/odrl/

ODRL-REQ (2004) Open Digital Rights Language (ODRL) Version 2.0 Requirements (Working Draft), ODRL Initiative, http://odrl.net/2.0/v2req.html, 24 November 2004.
Guth, S. and Iannella, R. (eds).

Stepanyan et al, *BlogForever: D2.2 Report: Weblog Data Model* (10/2011)

Wiles R. (2012) "What are Qualitative Research Ethics?", Bloomsbury Academic. ttp://dx.doi.org/10.5040/9781849666558

Appendix A. List of Tools for Digital Rights Management

The following lists some useful tools in developing digital rights policies. It is not meant to be an exhaustive list. It is intended to be used as a guide to different types of tools there are (indicated by the headings), and as a starting point for searching for solutions.

Guides

Advanced distributed Learning
http://www.adlnet.gov/
Guide on Re-use of Learning Material
http://www.reusablelearning.org/
JISC/SCA Intellectual Property Rights Toolkit
http://www.jisc.ac.uk/publications/programmerelated/2009/scaiprtoolkit

License

Creative Common License
http://creativecommons.org/
General Public License
http://www.gnu.org/licenses/gpl.html

Rights expression language

Creative Common Rights Expression Language (ccREL)
http://wiki.creativecommons.org/CC_REL
Open Digital Rights Language (ODRL)
http://www.w3.org/community/odrl/

Registries

Registered Commons
http://registeredcommons.org
Rights Metadata for Open archiving (RoMEO)
http://www.sherpa.ac.uk/romeo/
Safecreative
http://www.safecreative.org

Repository risk management

Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)
http://www.repositoryaudit.eu/

Repository and Content Models

Article on the DELOS Digital Library Reference Model
http://www.dlib.org/dlib/march07/castelli/03castelli.html
DL.org Digital Library Reference Model
http://www.dlorg.eu/index.php/outcomes/reference-model
Reference Model for an Open Archival Information System (2003 version – links to 2012 version)
http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683
Sharable Content Object Reference Model
http://scorm.com/

Rights Metadata Schemas

MARCXML
http://www.loc.gov/standards/marcxml/
Dublin Core
http://dublincore.org/
Qualified Dublin Core
http://www2.archivists.org/standards/qualified-dublin-core-qdc
METSRights
http://www.loc.gov/standards/mets/mets-extenders.html
http://www.loc.gov/standards/rights/
PREMIS
http://www.loc.gov/standards/premis/

Resources for Legislation

Cornell University Resource US code for Copyright
http://www.law.cornell.edu/uscode/text/17
Intellectual Property, Europa – Summaries of EU Legislation
http://europa.eu/legislation_summaries/internal_market/businesses/intellectual_property/
European Union Law
http://eur-lex.europa.eu/en/index.htm
Intellectual Property Office, New Zealand
http://www.iponz.govt.nz/cms
IP Australia (government site)
http://www.ipaustralia.gov.au/
LexisNexis (Search tool for US Law)
http://www.lexisnexis.com/en-us/home.page
World Intellectual Property Organisation
http://www.wipo.int/portal/index.html.en

Organisations
(see Sections 2 and 3 for privacy policies and rights management policies)

British Library
http://www.bl.uk/
Digital Public Library of America
http://dp.la/
Harvard's Web Archiving Collection Service
http://wax.lib.harvard.edu/collections/home.do
Internet Archive
http://www.archive.org
Internet Memory Foundation (used to be the European Archive http://www.europarchive.org/)
http://internetmemory.org/en/
PANDORA, Australia's Web Archive
http://pandora.nla.gov.au/
UK National Archive
http://www.nationalarchives.gov.uk/
Web Archive at the Library of Congress
http://www.loc.gov/webarchiving/
Web Archive, Bentley History Library, University of Michigan

http://bentley.umich.edu/dchome/webarchives/

Appendix B. Transcript for Interviews

The following are interview transcripts for interviews presented in Section 6. In the following, the boldface indicates the interviewer's questions, followed by the interviewee's response.

Transcript for "Paul Johnson"

Hello?

Hi.

Hi there. How are you?

I'm fine. Let me just close the door.

Thank you very much for agreeing to talk to me.

No problem.

Your name comes up in a lot of the publications that I have been reading.

I am trouble cause probably…

No, no, it's wonderful and it's very much in line with what we've been talking about…

I'm actually on this one…I have two different cameras. One is on the bottom and one is somewhere else – I'm never sure which one I am actually using.

Oh ok. Well, I see you now. That's alright…Let's see…I am not sure if you were able, if you had time to check out the link I sent you about the project…do you want me to explain to you a little bit about what we're doing?

I had a brief run through it just to simply see who you were. I get quite a lot of things through and I go off and check to see who I'm talking to.

Sure, sure.

So I've got some idea what the project's about. Yeah.

So essentially, you know, we know that a lot of work has already been done in this area, just in general in terms of digital archiving…and so what we want to do is really kind of focus a little bit on, let's say, some of the up and coming issues, new legislation, things like that…and then also things that could be particularly important for blogs. I'm sure you can imagine what these are: Third party content issues…

Yeah, yeah.

So with my colleagues, we drafted kind of a set of questions that I wanted to chat to you about. There's about 20 of them. We can see how far we get through and then, if the time starts stretching out a little bit too long then I can also send you the remaining questions and you can decide if you wanna have another chat about it or if you wanna drop me a line.

Ok, whatever suits.

Ok. Perfect. One of the first things we were talking about in the project is…you know, there's a lot of things talking about risk assessment but we wanted to know a little bit about what those risks actually are. I am not sure if you know anything about how frequently rights holders bring legal action against public institutions in the [*name of European country where the research subject works*] or abroad that are preserving digital materials. Or what sort of things those legal actions resolved in.

No, not so many. I mean, the kind of risks…we have seen people sue. But it tended to be sort of newspapers…there's a case in the UK, the Baranovsky case, and they sued The Times, I think, it was over the web archive of newspaper articles. And The Times have taken down or they published their correction and some such with to do with the actual paper version of it. But they hadn't actually made a similar note with the archived web version. So there was that issue first of all. The second issue that came up in that case was, under UK law, technically you have a year to sue for liable, a year from the first day of publication. And what The Times was claiming was that Mr. Baranovsky was out of time. It was after a year. What the court held was that every time somebody accessed the web page it was a new publication and the clock started running again.

Aha.

So if you look in the new UK defamation bill, there is now a proposal to change that…there also provisions…let me just see if I can…I cannot remember exactly what they are…the bill is still going through parliament…if you hear the bell that's just our bell tower. [*portion of transcript related to location removed*]

Ok.

[types something] Right…defamation bill…basically, there's a fairly major shake up of defamation law going on in the UK. That's the idea. That's what this bill will do. And there are changes with regard to the law defamation with regard to operators of websites. So things that are…so people who are hosting blogs may have more defenses than they had before. And I am pretty sure they also changed the time period. [looks something up] Yes, single publication rule…so essentially, any subsequent publication of the same thing – so somebody comes to a web page – it wouldn't then count as a new publication. It would be the same publication again. So if I put something on the web that was defamatory about you and you let the year go past you wouldn't be able to sue me even if somebody accessed it just yesterday.

Ok. And what if it's republished on another…?

If it's republished in a materially different form then my guess is that if you published in another place, that would be a materially different form because it wouldn't be on the blog…then the thing starts again, the clock starts again.

Ok.

But you can find the defamation bill. If you type in "defamation bill UK" it will pull you up the bill as it stands at the moment going through parliament.

Ok. And do you think in your…based on your experience and what you've kind of been keeping track on…do you think that most of legal actions are resulting in real financial damages or just a demand of the content be taken down?

The kind of people…I mean, I deal quite a lot with sort of higher education and archives and galleries and museums and [*name of a specific archive*] and those kinds of things. And a lot of the time, what the people are looking for is the stuff to be taken down rather than significant damages.

But there's always that threat, of course. So I can't say off the top of my hat beyond the Baranovsky case that I can think of anything specific. But I mean, basically, what you are going to be looking at mainly for risks is going to be copyright infringement, defamation…those are going to be your key risks, I would think.

Ok. That's actually what one of our next questions was…We were wondering what kind of legal issues that, you know, maybe [*name of archive removed*] and other web archives you've worked with what they experience. We were thinking of rights, statutes, licenses, surveillance, privacy, data protection…these types of things.

Yeah, the data protection is a good point. The problem with blogs, of course, is if you're doing a sort of web crawling to gather in blogs you run the risk of...if you're making this stuff available then the defamation point, the copyright point, data privacy…and data privacy is a bit unclear at the moment. I mean we have the ruling of the European court of justice in the Lindqvist case that says that publication of sensitive personal data even on your own website or your blog is publication. You're a data controller for the purpose of at least the Swedish legislation. I don't think we've had any similar kind of similar case in the [specific country removed]. I suspect we wouldn't…it's not something the Information Commissioner in the [specific country removed] would be particularly inclined to chase up because they'd be chasing everybody in the country who has a web page, basically…

Right.

So certainly copyrights, defamation, certainly data protection – those are people's main worries, I think. This is certainly the area where you are copyrighting in particular where you are likely to, at the very least get some sort of notice and take down.

Right. One of the other things we were discussing is…you know, this is a software platform, different bodies, different institutions might use it differently from others and we were thinking in terms of the opt in and opt out principle applied by the internet archive…because you know, there would be the idea of not inserting, not ingesting anything unless it's being suggested by the owner directly, directly input to the repository. And then there is this idea of spidering around and sort of collecting things. Have you…are you aware of any notable problems or risks that are specific to that?

I've certainly…I remember talking about the internet archive with [*name of archive*] a number of years ago. Because their point was "Well, the internet archive just spiders all that stuff" and I said "Well, the internet archive is essentially a massive copyright violation". And what you find is the internet archive…if you complain about it they'll take it down. A nice example of that was the debates over the Church of Scientology. There is very little of that material online because the Church of Scientology considers a lot of what it publishes to be its copyright works and will simply send notice to take down. The internet archive doesn't have the money to fight or the inclination or the time to deal with these things. So with a lot of these kind of things they'll just simply get taken down. And that was one of the problems essentially…people would say "Well, the internet archive collects everything." I said "Yeah, but even if it collects everything chances are you're going to have holes in the archive where people have complained." And unless you are in a position to make a determination about the copyright yourself you're going to do the same thing. And the [*name of specific archive*] at the time basically had more or less decided they weren't going to collect everything. They were going to collect certain specified websites and they were going to negotiate with the owners of the websites in advance over issues such as "Do you own all the rights to the material of the website?", "Can we archive it?" and "What can we do it with it once we've archived it?" And that's really the big question with this. It's not so much the archive. A lot of people don't mind you archiving stuff. It's about what is it going to get used for.

Right.

To the extent to which third parties might take it…again, it depends on what the third parties are going to do. If they are going to use it for research – and this is what we were talking with the [*name of specific archive]* which is lively interested in archiving the [*country specific*] web so that future researchers could look back and actually get some idea what was going on in the [*European country*] in terms of blogs and websites and so on. And so the idea was that this stuff would be available to potential researchers. And even then, the [*name of specific archive removed*] was saying: "Well, we're going to probably hedge that around a bit. It might be that you have to come to the [*name of specific archive*] to look at it. We're not going to make it available over the web. You're actually going to have to be physically on the premises to look at it." I suppose you could set up a system whereby you'd be an official [name of specific archive removed] user or some sort and maybe be able to access it online. But in general, I think the idea was very much it would be a resource for researchers accessible at the [name of specific archive removed] and for research purpose and not commercial use. And that's a difficult question right there.

Do you think…obviously, both of these strategies can lead to spotty collections requiring…which do you think…I mean if you had to imagine which approach, let's say, has the least risk associated with it in terms of having a spotty collection. Seeking permission or kind of what the internet archive is doing, just you know, taking down…or maybe a combination of the two, I am not sure how you see that.

I mean…there is another risk as well. It's also the risk that you are collecting material that is illegal in itself, like child pornography, extreme pornography in the [specific country removed] which is essentially pictures of people having sex with dead bodies or animals or pictures that look like they are doing it. And it's illegal to have these things. It's a strict liability offence in the [specific country removed] if you actually possess them whereas obscene materials, sort of bog standard pornography and the like, it's only an offence if you're distributing it. But if you got an archive that you're making open to the public you'll be distributing stuff as well. So…

Has that ever happened to the internet archive?

Not to the internet archive as far as I know. The US, of course, is going to have a different approach to this because in the US, you have the first amendment so it's harder to ban stuff in the US than it is in the [specific country removed]. Again, what I am saying is these are risks. I had to get this issue with the [name of specific archive removed]. The [name of specific archive removed] says: "What's the risk?" and I said: "I really don't know." It's almost impossible to quantify the risk. It's impossible to quantify the risk in terms of copyright because it depends on the rights holder. You may get rights holders who would like the material taken down but they are not going to take it any further. You may get rights holders who want to make an example of you. And individuals are becoming more aware of…if not the potential value, the fact that other people may be using their material and they want credit for it or whatever. And there's an interesting article on, I think it was the BBC website - it might have been BBC, it might have been The Register - about people complaining about people using their Instagram...so people are complaining [00:15:08 inaudible] instagram photographs and e.g. putting them on t shirts or whatever…and it's a copyright infringement at the end of the day. It's pictures of an original picture. And you get professional photographers coming on and say: "Well, we will take legal action. We will send you a letter saying we'll take legal action if you don't take it down or whatever else is required or we'll sue you." So I think there's a growing perception amongst the general public that – when you put things online, they are putting things online for a purpose…maybe to show their friends…but they are not necessarily putting things online to be widely shared, to be archived or in particular, to be reused in any form without their permission. And this is one of the things I was saying in the Report that I wrote for the [name of specific archive removed]. One of the things that the [name of specific archive removed] should perhaps be considering is how you make the archiving part of the interaction with the general public. So what I was suggesting was that they could team with Facebook or they could team with any number of social networking sites. And what you actually do is you offer the public the

archiving options. You say "The [name of specific archive removed] is interested in archiving these things, if you would like to do this tick the box." Or if you're really cunning you say "If you don't want us to do this tick the box" because people are much less likely to tick the box. So what I was suggesting for the [name of specific archive removed] at the time is that the general public are getting more savvy about this. They are starting to get a little bit more in tune with what intellectual property is, how they personally could potentially use it. The risks of an intellectual property suit are…I would suggest, unless you've taken material that is owned by a professional photographer or professional publishing house or something like that…the risk is actually relatively small because bringing a copyright action is an expensive proposition. So a person would have to be fairly sure they were going to recuperate whatever money they have invested in legal action. And the average person in the street is not going to do that. On the other hand, it's probably not the best way to build a relationship with the general public…

That's…yeah, I can definitely understand what you're saying there. This was part of this sort of social aspect of this project. You know, how to actually engage the public in the concept of preservation as such. And that's definitely not alienating them, We definitely don't want to do that.

That's right. I mean I've said the digitalized project which was…I mean, the more recent one was called "digital preservation", more generally. But the digitalizing really was about archiving personal digital archives by which we essentially meant blogs, Facebook pages and so on. And the easiest way to get around that, to get around some of the legal issues at least is to invite the public to help you do the job. One of the other things I was saying was getting the public to add metadata to the information that's being uploaded…because that then cuts down the immense amount of work the archive has to do.

Sure.

Again, the problem with ingesting anything that comes from the public is that the public have a rather – how shall we put it? – different attitude towards intellectual property than the businesses do. So you will almost certainly – if you ingest without some kind of filtering you're going to be ingesting stuff that is other people's copyright.

This was one of the things we were actually thinking about even in terms of…'cause this is an international project. It's EU funded and we are thinking even about multi-jurisdiction rights on the preservation of content. What are the implications of that? Or when you're dealing with…Let's say, a blog from [a European Country] contains material that infringes rights in France or the US…what do you think are the implications of that?

Well, in copyright terms, a copyrights holder can sue in any jurisdiction that recognizes that copyright. The whole purpose of having the Berne convention is so that rights holders' rights are recognized in other jurisdictions. So in principle again if you're using people's…let's say you've ingested a website from the UK, it contains copyrighted works from a French artist or French photographer. You're infringing their copyright in any country that recognizes that copyright. So if your servers are based in France or Germany or wherever, they can probably sue you there. In fact, they could almost certainly sue you in their own jurisdiction. The question then is whether the courts of the jurisdiction you're in will enforce a judgment of the French court.

Right.

If you're with me.

I think I am. You know we were specifically talking about the issue of France. If somebody…also how you actually tell under which jurisdiction…we were wondering is this related to domain name? Or if a person in the UK used a public blogging platform e.g.

Wordpress and they posted material to the web…and the work is from a French author…to which jurisdiction does their material belong? Is it the domain or is it their person? Do you know what I mean?

Well, if you have a work by French author – the French author is living and working in France – they will have a copyright in that work in France. But it will be enforceable in other jurisdictions. If I infringe the French author's copyright in the UK I can be sued under UK law because UK law recognizes the copyright of French authors because of the relationship between the UK and France, and any other country. So there are…again, what it comes down to is this something likely to sue for that? And again, it depends…you get people who are very litigious. If you try copying anything that has Star Wars in it you'll find out that Lucas Films are extremely litigious. They will sue you. Or they will threaten you with a lawsuit.

So then…what do you think…so what I gather from what you're saying is that your recommended strategy would be to involve the public in actually what content gets ingested and how it…

My starting point would be: How risk averse are you? The internet archive is clearly not that risk averse. And they come out with a strategy which is essentially "notice and take down" and they are essentially betting nobody is going to sue them as long as they take the stuff down reasonably quickly. The only people who can tell you how effective that is is the Internet Archive. It does lead to you having a potentially a patchy archive and there is some e-material missing from the internet archive. Again, the extent of that missing material would be something that would be more familiar with than I am. They'll be able to tell you how many take down requests they get on an average basis and that would give you an idea of the cost of running an archiving platform. The more risk averse you are – the [name of specific archive removed] e.g. were quite risk averse because they said "We're a national authority – we really can't be seen to be going around downloading stuff willy-nilly and breaching copyright" and that is why they adopted the strategy they did, which was essentially saying "We're going to pick". We're going to essentially curate. We'll decide what gets archived. We're going to ask for the necessary permissions and we'll work that through. And my understanding was that was taking them longer than they had expected them to take…because archiving websites…the websites may contain all kinds of different copyrighted material owned by different people. And getting the rights could be quite complicated. That was what sort of led me to make the suggestion that I did with regard to archiving personal digital archives which was to say "Well, it seems the most sensible approach is to build that into the public experience of having a personal digital archive." Basically, what you're saying to the public: "We're offering you a service." Then in return for that service, we get access or future researchers get access to it. So in many ways, you're making a PR pitch. You're saying to the public: "You are important enough that we would like to know".

Sure. Under this idea you had how did you imagine it working with…you know, when people post videos on Facebook. That's a lot of the content on Facebook – who would be responsible in that case for…I mean, what would have been your recommendation about that? Who would have been ultimately responsible for those types of copyright issues?

Essentially what you would…well, this is what it came down to…perhaps you could involve the public in providing some metadata. The original position I got from people in the business "The public don't do metadata." They won't do it. And I said: "But they tag things." People tag things on Facebook all the time. Or if you give them the opportunity to tag things and you make it clear that the better tagged things are the more likely their work is to be preserved for prosperity. Again, you're providing them with an incentive to do it. Now, in that circumstance you might say…I mean, when the member of the public…I mean, this is all sort of hypothetic. You could have a…when they first set up their web page you could have a box they can check that say "I would like this archived" or you could allow them to selectively archive over time. And if you allow them to selectively archive over time you can ask questions about where did you get the information or who

is the owner of this? Again, it's making it as uncomplicated as possible. Essentially along the lines "Did you take this video?", "Did somebody else take this video?" And you might then say "Do you have the permission to use this?" You don't want to make it more complicated than that. And then at least, if you are then approached by a third party who say "You're using this video without our permission" you can say "Well, we'll take it down immediately. The reason it was up there was that the person who put it there said it was theirs." So you're giving yourself a little line of defense. It's not bullet proof but it's distancing yourself a bit from actually getting sued.

Moving into a more perhaps general area…we were wondering also about contract bond in the form of software and content licensing. How that sort of reshaping the digital landscape, you know, what kind of impact that has on the preservation and curation of digital content by public organizations.

Well, some of that was covered in the digital preservation paper as I looked at software and the preservation and the emulation software because…

Right. That's a big issue…

…what could be an issue for archiving down the road is what happens when the software changes? Can we emulate it? And the law, certainly the law in the UK, is not terrifically clear on that point. So we have a situation where we think you can emulate the software as long as you don't actually copy the software. So I mean you got the right copy as an archive in theory for preservation but those don't apply to digital works as yet. The archiving for preservation purposes in the UK basically applies to tangible items. And that's one of the problems in the UK. We press the archive's act, the Legal Deposit Act 2000 which was supposed to allow the Secretary of State to extend what could be archived and preserved. Nothing has really happened since. The legislation is there but we're not really advancing it. So we have the situation where things like software, films and the like are not really covered by clear law in the UK in terms of "Can we go back copying them for archive purposes, for back up?" And what can we do with those? Can we actually make them available to the public? Under what circumstances can we make them available? Can we place them on the internet? Can we place them on a computer in a library? Or can we not do anything with them at all until copyright expires? Do we have a dark archive?

Right. I mean what do you think are the best practices that currently exist in navigating this big grey area?

It was quite interesting looking at the BBC's Domesday Project. I did some work on that many years ago when we were first looking at it. Are you familiar with the…?

Yes.

Yes. And the BBC have released some of this material. I think one of the two discs is actually available on the internet. You can actually look at it and search through it. The other part of the disc is not there because they simply cannot be sure about the clearances. And part of that was simply because they didn't have any [inaudible 31:31]. They either didn't have it or they didn't keep the necessary documentation to demonstrate that they actually asked for rights and they asked the right people for the rights and so on and so forth. So again, the BBC is relatively risk averse. In the sense of they can't really be seen as a national broadcaster to be willfully breaching copyright. So...it's debatable to whether we will ever see the full Domesday Project online. There have been attempts to produce commercial versions of it. As far as I know none of those have actually happened here. And it is very much down to sorting down the rights. Again, it depends how risk averse you are. And it's essentially going to be my answer to a lot of these things. There are methods you can use to reduce your risk and if you can reduce your risk to a level you're comfortable with as an organization then you're going to be able to archive more material. If you are risk averse, in the sense of you don't want to – or you cannot for reputational purposes - be seen to be infringing

copyright or being less than diligent about dealing with it properly, well, then you're going to archive less.

So do you think then that for a project such as ours where essentially…you know, we would like to be able to provide some sort of recommended policies. Thinking about the fact that there could be different types of institutions using this platform that part of our policy development should be to suggest to our clients to assess their own level of risk aversion?

Yes. I think that's right. The interesting thing is that over time for instance the [name of specific archive removed] has become a little less averse. I mean I wrote a report on web archiving back in about 2000. And what I was asked to do was to write what the risks actually were. And I did write what the risks were but I didn't really go to the same extent as I did in the digitalized project how you might handle those risks and the issue of risk aversion. And I terrified the [name of specific archive removed] basically. The head of digital preservation, he said he had the report on his desk and he said he used to flip though it on a regular basis…and that wasn't what I set out to do. I was sort of setting out to say: "Well, ok, these are the legal issues." But that was the brief I was given. And I think over time, we have become aware, we have become more comfortable with the idea that we cannot archive quite a lot of things without people complaining. As long as people think that they are being treated fairly…and certainly, if you are looking at archiving blogs and things like Facebook and what have you – as long as people don't think they are being taken advantage of – and that's the ordinary person in the street – they don't seem to mind too much. You are always going to have difficulties with the professionals because if you're a professional photographer you make your money out of selling works and it's damaging to you to have your works given away for free or available for free. And those people are always going to complain if they know about it. And what you have to do for those kind of people is to provide a mechanism by which you can deal with those complaints as rapidly as possible. Essentially, when it comes to assessing to damages for intellectual property one of the things the court will look at is how long was the material available, how many people downloaded it or could have downloaded. What's the damage to the rights of them? If you dealt with the complaint quickly you can perhaps demonstrate that not that many people have downloaded it, you acted as rapidly as possible to take it down. In those circumstances, I think, by and large you're not going to run into too many difficulties. I mean, there is also the issue, of course, of the e-commerce regulations and the availability of that sort of limited defense for interred intermediaries with regard to your knowledge of infringing materials. The problem is if you are archiving…if you're basically saying "We're ingesting everything" arguably you know you are going to be ingesting things that are infringing because you're deliberately going to have to do that. It's not like people are coming to your site and depositing things that you don't know are infringing but might be. You're saying "We're just grabbing everything". So that causes problems with that [inaudible 36:50].

That's definitely something we were thinking about also. Just what you said that the act of record keeping in terms of access of various materials could be important then in the case of a litigation.

Oh yeah. If you can demonstrate that the material…the problem, of course, even in those circumstances is that when you're dealing digital artifacts – one download from you can still mean a 100,000 iterations of the thing once somebody else has got it. But it's a form of damaged reputation. You'll say "Ok, we made it available. We probably shouldn't have done. But only one or two or three people downloaded it. We took it down at the moment we were told that there was an issue." In those circumstances, whether you're dealing with copyright or indeed with defamation, you're limiting your liability. And again, that brings you back to the question of your risk aversion. Can you live in your liability through those mechanisms to something you're comfortable with? Or more often, something your insurance is comfortable with?

Yes. This is maybe a specific question…I am not sure if you'll know the answer to it…but can blogs that use a variety of materials, you know sources from different content providers, a

blog that discusses a film and uses clips…or some of this ever be considered to be an orphan work

Oh…orphan work…well, if you are using a part of a copyright work for the purpose of criticism – certainly in the UK, that may very fall under fair dealing. And there is a big difference between US fair use and UK fair dealing. The US fair use is a much broader category of things and the UK courts interpret fair dealing very narrowly. So for some things you might be able to…and certainly, somebody running a blog if they are criticizing a movie, if they are criticizing a book, you might be able to use excerpts. But again, a court would look at that and say "Is the amount of the book or the film or the poem fair?" Is it proportionate to the purpose you're using it for? So putting half of the film in your blog for the purpose of criticism is probably going to be seen as excessive. And the issue of orphan works is up in the air a bit at the moment because the easy use in the process of working at it directly [00:39:53 inaudible] orphaned works. I think there is quite a lot of pressure in the EU and in the UK for a greater ability to reuse orphan works. Certainly the [00:40:13 inaudible] reports, the report commission by the government a couple of years ago, suggested being able to reuse orphan works would actually be quite valuable. There is a lot of material sitting around that simply isn't being utilized but could be utilized. There is a lot of resistance to the idea of orphan works from traditional publishers, from photographers again and in particular, professional archives because they are concerned at the ease with which something might get declared an orphan work.

Sure.

And whether or not they would be able to successfully defend their intellectual property in those circumstances. So orphan works is going to be a complicated and controversial area, regardless of whether or not we get a directive, simply as we shake how where the boundaries...you know how we define what an orphan work is, what happens when somebody turns up and claims an orphan work. In other words, it isn't in fact an orphan work…will there be no liability for somebody who's used an orphan work in the honest belief that it is an orphan work? Will they have to pay a license fee for the use of it? Prior to that disclosure that it's not an orphan work or only after it's been declared not to be an orphan work? All of this is up in the air. And if it's…I mean, the big problem with dealing with the general public and intellectual property is the general public have a general idea of what intellectual property is. Most people are aware that downloading a brand new music track from the internet is probably breaching copyright. Whether that stops doing it or not is another matter. They are also aware the uploading it or putting it into a torrent is breaching copyright. And that you are more likely to be chased after if you are an uploader than the downloader. So the general public have a vague idea about that. The general public is very very hazy on the issue of what they can do with material unless it's already on the internet. Academic are very very hazy. A lot of people tend to assume that if something is already on the internet they can just use it because it's already out there. And so the other thing I have tended to say when it comes to dealing with the public is: If you're going to do these kind of things a strand of your work is going to have to be educating the public as to…you know if you want to deposit…or if you want us to archive your material, these are the ground rules. These are the things we really can't accept. And again, that depends on the way in which you are ingesting the material. If you're just going around and grab it…well, you don't have that degree of ability to educate the public, to talk to the public. If you are working through a third party like Facebook you at least have the opportunity to reach out to the general public and say: "These are the rules as regard copyright, defamation or whatever. These are the risks that you run yourself as an individual putting this material. And this is why it is a bad idea for us to archive this or why we can't archive this if you have that kind of material."

Sure. I mean, what do you think…just in terms of what's coming in the future with respect to rights management…

Futurology.

Futurology. A little bit of futurology. What are the kind of main changes that you think are

coming? You know with this landscape changing…if the public can get more involved. What do you see kind of coming?

I think…well, let's start from the beginning. The technology for tracking and tracing copyright infringement, defamatory statements, breaches of data privacy is only going to get better. Which means that if you are hosting material that is a copyright infringement, defamatory or breaches data privacy you're much more likely to be called out in the future than you are now. Which means you're going to have processes and places to deal with it. The general public is going to be a bit more aware of these things but nowhere near enough aware. I think for the short term- the short to medium term for you to be a 100% reliable - the fact that they know what they are talking about when they say that they've got the rights to the work and you can archive it. But you can use their…but you could harness the general public. And I think the general public are going to be more in tune with the idea of things like metadata. We wouldn't even need to be talking about metadata. What we'd talk about is tagging. We use the terminology of the public. The public is already familiar with tagging things. Tagging things with a little bit more data such that their webpage, their blog, their Facebook page can be archived is not going to be a huge amount of work. We can sell that to the general public as "If you do these things your personal digital archive is going to be preserved for posterity" – that's the way I would sell it to the public. And I think the public is quite keen on that. I mean you only have to look at the growth of the genealogy websites. And I mean my dad does this. My dad's been busily building the family tree… I get an alert every other day. You know such and such has been added to this. Again, if you could hook in to that kind of desire for doing something for posterity – that's a very powerful tool. It's not going to get you everything but potentially, it could be quite a powerful tool for gathering, at the very least, a picture of what the people are doing with social networking tools, with the internet, with the web at a given point in time. I don't think we've ever been in the position more than a partial snapshot. And I don't think we're ever really going to have the ability to have more than a partial snapshot of the internet. I don't think it's realistic to think that we are going to get everything. And it's also arguable whether or not you need everything. I mean the internet archive archives a lot of stuff which quite frankly is junk. As the technology progresses the other thing we're going to see is digital archives, people doing digital preservation...beginning to develop tools for better curation. In other words, enabling us to decide what we're going to keep for posterity and what we're not. And I think that will be true even if we get to the point with storage technology where we can keep everything. And I am being forever told by people that we're almost there, we're almost to the point where we could keep everything…the question then is: Do we really want to keep everything? And even if we can keep everything will we ever be able to make sense of it all? I have always tended to look at the sort of preservational archiving sites as being – at the end of the day as the technology matures – as being a curation and preservation. And therefore there will be some discrimination in terms of what we keep and what we don't. What's useful and what isn't.

Who do you think will be making those decisions ultimately? Do you see that as being like public bodies and institutions, individuals…?

At the moment, I see it being very much the people who have traditionally done that. But again, there's something that lends itself quite nicely to that whole sort of crowd sourcing process as well. Getting other people engaged in archiving, preservation…getting other people engaged in that production of metadata, that means of making information, making blogs, making Facebook pages useful. So a lot of what I envisioned in the digital lives sort of sense was much more use of the public or much more interaction with the public. Not directly necessarily, but through third parties like these social network services. And the hardest thing of the general public's interesting in producing thing that are valuable or that are heritage kind of…you're selling it as heritage.

Yeah, I definitely think…we've talked about that a lot in the project…

I simply can't see in this…particularly in these kind of economic times…I can't see that, you know, even people like the [name of specific archive removed], national archives, are going to have the

resources, are going to have the money to deal with the influx of data you could be getting if you really wanted to.

Well, let's…[interviewee still talks] oh, sorry. I'm sorry.

Go ahead.

I was just saying you just decided two important things, you know, if we're thinking about the top challenges that you see for the digital preservation community, in particularly with respect to digital rights management policy…I mean, obviously money is a big issue. Resources are a big issue. What do you see as being the main challenges that we're facing?

As we said, we may get to a point where we can archive pretty much everything. But the challenge then is producing the tools to make that everything useful. And to the extent that you can alternate that, well, that will reduce some of the costs…some of the drain or the use of resources. But as I said, I still think if we are going to have a successful digital preservation strategy it has to involve the public, particularly if you're looking at the person in digital archives, if we're looking at the archiving of traditional or even digital works in the sense of things that are produced by commercial entities or whatever. Well, we can do that without the public. If we're wanting to preserve personal digital archives the public have to be involved. And the challenge there is how do we get them involved? How do we get them interested in the things that make the information valuable? How do we get them to help us sort out the material that is interesting? How do we get them to sort out the material that is infringing? How do we get them to help us sort out the kind of things that they want to see? Because that's going to play into it. And a lot of this preservation and archiving at the moment seems to be going on without a lot of public influence. There is a lot of "We'd like to collect the web" or whatever…but the public…the need to archive has made some steps along the way with the wayback machines on. But I don't certainly get the sense, sitting in the UK, that there's been a great deal of engagement with the general public in that kind of education. Why are we interested in keeping your blog? Why are we interested in preserving your Facebook page? How can you help us do that? And why might it be interesting? Why might we want your blog saved forever? What conditions do you want to put others preserving your blog? We know that general public are – some members of the general public really don't care, everything could go. But also know a lot of people have – even though they are putting their material on quite often what are essentially public access sites – actually still have very clear ideas of how don't want that material to be used. And again, this engaging with that in terms of working out where points of resistance are likely to be. And how you address those points of resistance. If people don't want or aren't making stuff available for you to archive why are they doing that? Is it something that can be overcome through better education? Is it something that could be overcome by giving them some kind of control of the material that you have archived? Again, it sometimes tends to be with the preservation and archives is the thing that one's it in, it's in. And it's ours now. That's not gonna necessarily play well with the public.

Actually, this…you know, I have just one more question for you. When we were thinking about a community driven rights management policy, is it a viable option – like the same way that Wikipedia's knowledge management is driven by public knowledge – do you think that something like that would be a viable option, especially when you're considering preservation of weblogs or other personal types of sites like that?

Good question. I think you have to think very carefully about what product you want to come out at the other side of this. Just in the same way as…you know, what you get out of Wikipedia is reliable to a point. I use Wikipedia for…certainly things to do with computer science, to do with the sciences. There, she tends to be pretty good. The stuff relating to individuals, for controversial issues, Wikipedia is somewhat less reliable.

Sure.

But again, to sort of feed that back into this whole idea of – and I am going to come back to risk assessment – That might be all you want. You want something that is relatively reliable. It's not going to do everything for you. There will always be a role for the professional curator or the person who oversees, the senior Wiki-Editor if you will. And I don't think any of us can really predict too far into the future how these things will develop. I mean we've been using Facebook pages, social networking sites, for under a decade. And we're sort of saying "How are we going to preserve these things into the future?" Well, we don't really know what they are going to look like in the future. We can handle it to some extent, e.g. by engaging with the people that are creating the material which is true on Facebook, places like that, to a very large extent the public themselves. The other side of this is over time is negotiating with rights holders in the same way that Google has had to negotiate with rights holders. It helps, of course, for Google's point of view that they've got billions of dollars…if we all had billions of dollars we'd be laughing. There would be no problem doing this. Again, a part of this is working out how you can do these kinds of things without making the rights holders feel threatened. The rights holders tend to feel threatened very very easily. Hence the kind of arguments around things like orphan works. So again, in the longer term…difficult to say how these things are going to pan out because what we're seeing over time is certain industries of these becoming a little bit more comfortable with dealing with their works in the digital environment. We're a long way off a situation where we can get around, we can avoid a legal liability simply by talking to the rights holders.

That's wonderful. I really appreciate the connection that you're making between educating the community and digital rights management. I think it's definitely one that we've been talking about a lot in the project. And also just in terms of what a good preservation strategy is to get…

It's quite interesting seeing how the public is gradually coming to terms with the idea that it's not necessarily a good thing to download music without paying for it. The public is starting to come to terms as the public are creating now more and more content. You do get some very interesting disjoints. I mean, the instagram example was a case in point. On the one hand, you have members of the public saying "Oh, we'll download this stuff from the music industry, we don't have to pay a penny for it." But the next moment, maybe not exactly the same people, but certain people from the same pool are saying "Hey, these people are ripping off our instagram!" There's been a couple of books…there's one particular book and I can't remember what it's called…but it was very much that idea what happens when the people who have been downloading this stuff for free get to an age when they are the producers of this stuff. That's going to be really interesting…All of a sudden you're switching from "Hey, this stuff's free" to "Hey, this is my stuff!"

Once you have stuff.

Yeah. And all of a sudden those intellectual property rights that we've been deriding for so long are suddenly quite an important thing. And I think that's the kind of transition we're going through. It's the kind of transition you have to deal with when you are looking at digital preservation, when you're looking at digital archiving. You have to understand we're going through a transition which we have not completed yet where the public are still coming to terms with what it means to own things, what it means to own digital material, owning rights in digital materials. And even the big players are coming to terms with "How is this going to affect our bottom line?" Whereas a few years ago, the music industry, the film industry were saying – as they frequently say – "It's going to put us out of business". They are finding ways to work around this. They are finding ways to work with people who are doing digital archiving/digital preservation. It's one of the things again I say in the digital preservation coalition paper is that nobody out there is saying we don't want to preserve this stuff. Nobody's saying it shouldn't be archived. What everybody is saying is "Yes, we'd like this to be preserved", particularly the content industries, "but we want to limit how people can reuse it". And it's the reuse that's the problem. The preservation part of it is less problematic. But we don't want to be preserving this stuff and sticking into the box for around 20 years…that's essentially

what you know…if we stick rigidly with copyright laws this is what we're at: We're at the author's life plus 70 years and dark archives. We can collect it but really, we're going to be wanting to use it pretty rapidly thereafter. And it's going to be awfully difficult to persuade government and the EU or whatever to fund stuff that is that speculative. "Oh, we'll collect this now and you'll be able to use in a 120 or 150 years…" "Yeah! We're going to pay for that!"

That's a very very good point.

Yeah.

Well.

Have I sort of answered your questions?

No, absolutely you have. And I am just sitting here thinking that I am sure when I listen to this again that I may have some additional questions for you. I am wondering if that's ok if I can contact you again.

Yes, that's fine for me. Are we just…

It's been a wonderfully helpful conversation that I have just had with you so I really really appreciate you taking your time. I know we've been on the phone quite a bit. So I just wanna be conscious of your time here.

No problem. Certainly, a lot of what we've been talking about is touched on in the reports that I wrote. So I mean this conversation in conjunction with those things…and I tend to be a bit more detailed in what I wrote…

I did have a look, you know, and I was especially looking at your recommendations. I kind of saw the direction the conversation could go into for sure. So I wanna take some more time to familiarize myself with what you've written, especially as regards some of the legal cases that you've referenced in, things like that. And after I have had a chance to listen to the interview again, maybe I'll be getting back in touch with you about more specific things. And so, of course, in the meantime, if you realize you have any burning questions about Blog Forever, what it is that we're doing you're certainly welcome to get in touch with me as well.

Fantastic.

It was very lovely talking to you.

And to you.

Interview Transcript for "Gary Fields"

[speaking while tape is turned on]

Of course. And if you'd like I could even send you a copy of the recording when we're finished.

Yes, if you can. It's not essential but if it's file size then send it, yes.

Thank you for agreeing to talk to me.
That's ok.

I'm not sure if you know my colleague [Name removed]

I know [name removed] very well. We're former colleagues.

That's right.

So I used to work with [inaudible].

He was the one who recommended to me that I speak with you. I'm not sure how much he told you about "BlogForever" as a project. Have you…?

He and I talked about…he came to talk to me about it and I think he ended up talking about blog [inaudible 00:47] guide him so this led to having this conversation, I think.

Right. That's exactly what it is. Basically, I am responsible for the deliverable that we're working on with this project that's related to rights management and issues of rights management and how that relates to collections management and other sorts of preservation activities. And [name removed] recommended to me that I speak with you. We have a list of some questions that have basically been following us throughout the project so maybe, it'd be good for me to start out with some of those questions and then we'll see where the conversation goes.

Ok.

And I'll try not to take up too much of your time.

Ok. Fire away.

Ok. One of the main things that we were talking about is actually to understand how frequently rights holders bring legal action against public institutions. And whether or not those legal actions actually result in damages or a demand of the content be taken down. I am not sure what your experience has been with the [name of specific archive removed] but we wanted to get a feel for the practical considerations involved.

Sure. Well, I would say this is probably…probably…Don't quote me on this. But in my experience which is as old as 6 months of work directly with the [name of specific archive removed] and then some acquaintance with [name of specific archive removed] for a couple of years and all that…my sense is actually a bit for the [name of specific archive removed] it is very infrequent. Very very infrequent indeed. That may be because it operates on the basis of explicit permissions in any case.

That's right. So it's not an opt out sort of policy…

Yes. Under normal circumstances it's definitely not the normal way of proceeding. So that's a big…on occasions there's still a theory of risk that actually a site owner who licenses us to archive their content doesn't have full rights of all of the third party [inaudible 02:59] what they have - which is possible – but hasn't presented himself so far. So the second part of the question is [inaudible] if I call for the answer because you don't have any test cases to work on.

Ok. And what are the kinds of legal issues that are typically experienced? Is it related to rights or statutes or surveillance? Privacy issues?

We provide…in making risk assessments, we provide for all of those things. My sense is actually it's different to different jurisdictions. So my understanding is…the Danish…I can't speak for all of them but anecdotally, the Danish experience is that their legal deposit web archive is available on site only. It's very very locked down, [inaudible] and very restricted. And I think the reasons for that

is they all have particular concerns about data protection. Whereas the aspect of it varies by registered [inaudible] some places there's less sensitivity around that has to do with its actual property, I suppose. But we would provide for all three. And so as I said we haven't been tested in order to have a really very full picture of what the main complaints are. We don't get enough [inaudible].

Sure. You just mentioned jurisdiction. We were thinking of the implications of multi-jurisdiction rights and the preservation of content. So for example when we are looking at blogs, what would happen if a blog from e.g. the UK contain material that infringe rights in France or the US or New Zealand…we're not exactly sure. First, how web archives determine, like what information is under their jurisdiction…

Sure.

…and what to do with multijurisdiction rights of other parts of content.

Well, I am not quite sure what the defamation of the stroke [inaudible 00:05:09] of implications are between jurisdictions…so what rights an individual in e.g. the US has in e.g. German law to seek regress for defamation within a website published within Germany. I don't quite know how that works to be honest. It's not something we've tested particularly. I suppose its [00:05:44 inaudible] property things are slightly more straightforward [inaudible]I guess that actually copyright is perhaps slightly more evenly spread, is more consistent across legislations. But perhaps the data protection and defamation restrictions... But I don't really know. But it's a good question. But I am not of much help there.

We were just thinking, you know, also for example in the case where a person in the UK uses a public blogging platform like Wordpress…whose jurisdiction does it belong to? Or if there's a blogger from France who's using a UK blogging platform, you know…I don't know if you've had any issues of this or if you have experience with that…

No, I haven't but I guess a lot depends on whatever the agreement between the platform and the author is. And I guess that probably varies from platform to platform. I guess. In print, there is an analogy with cases where publishers in fact own…or not actually the publisher, there's some sort of learning society that retains the title of publisher in the content but the publisher simply prints and distributes. And I suppose that may be the case. You know, I myself have a Wordpress blog and I'm based in the [name of specific country removed] and the Wordpress.com thing isn't and I'd have to go back to the small print, too, to figure out who it is that probably is the publisher of that material. I've always understood intuitively that it was me rather than Wordpress. So the publication that occurs is in the hands of the author. I would guess. But again, I don't really know where that's attested anywhere. It's an interesting question again.

Kind of thinking about in terms of preservation activities as a web archivist…I'm sure this has come up in your work as well…where do you see the complication between certain types of preservation activities and respect of digital rights?

That, I guess, would be a question…you'd be better off talking to my colleagues who are more embedded in the actual preservation work. My role is engagement, ways of communication and that sort of thing and [inaudible 00:08:22] that sort of thing. Our preservation approach is by and large more with emulation rather than migration. And we'd actually build the content…It's work based replayed through the way back machine and so the preservation work takes place within way back and changes would be up and into the archive, too. And so I…what sort of rights issues…what's your kind of theoretical case in this?

Well, I think one of the issues is of making copies.

Sure.

To what extent…and also the format that we present the blog. If it's not presented in the exact in the exact format in which it can be found on the internet. That would be one particular issue. I am not sure if you've thought about that. Do you have a standard format that your materials are presented in or is it…right, you said that – exactly, "emulation rather than migration"… So that's one of our issues. That's one of the main issues. Also preserving various types of content and what form…

Sure. I mean, by and large we capture a blog of a site and play it back as well as we can, given the limitations of both the calling and then the way [inaudible 00:09:58] at the the time. And so the issue of making a copy of it is covered by the explicit permission that we have from web or site owners. And then we don't subsequently…I don't think we end up making more copies of the same two although I think it's possible. I'll have to go back and look at the text itself. I suppose it's possible that we might have provided for doing that if we needed to. I suspect that a wise drafter of such an agreement would do so. And then as to whether actually it is not reproducing the exact replaying a site in a way that differs subtly but still clearly from a live version. I suppose that couldn't really raise some objection. But I guess that it needs to…again, that can be met in terms that there is an agreement with the depositing site owner that would say something along the lines of "we do our best to replay this in the way it looks like it would [inaudible 00:11:09] subject to the unavoidable limitations of available technology etc. etc."

But you advise against e.g. like an own look and feel of an archive in which case the actual structure of the blog is not preserved? Let's say it is preserved but it's not necessarily the first.

I wouldn't necessarily advise against it but it would just be…well, you know there are pros and cons. If in effect…[00:11:52 inaudible] what you sort of looking at in BlogForever is really actually not a web archive in the sense that you have something that's replayed in its existing look and feel it's really as if it were digested. And you have a database blog content as if it were presented through a single viewer, within a single visual identity.

Right.

And that, it seems, if that is something that is agreed to by the person giving whose content it is in the first place. But I wouldn't say that one approach was necessarily better or worse than the other that is different. It's just a question of making sure there's clarity and agreement from the people whose content it is to begin with.

Have you come to any specific challenges with the preservation of web blogs with regards to rights management? Have you experienced personally anything to this extent?

Not personally. But again that is partly because I am not at the center of the preservation or bridge. I can send you in the direction of colleagues who are more likely to have done…my sense of it in many ways from a capture and replay point of view, many blogs are quite straightforward. They don't depend on much in the way of particular plugins and flash or media [inaudible 00:13:24] from the part of [inaudible] so any of that stuff…really, they are relatively simple and regularly structured. And the data is structured with recognizable fields that describe what's going on. I would guess in some ways actually a blog's owner runs straightforward from a preservation point of view. But then I'm not a preservation specialist, so…

Yeah, it would be wonderful if you could direct toward anyone working with you that I can maybe talk to about some of the preservation activities in the relationship with rights management. That would be wonderful.

Ok.

Again, I am not sure if this is something that you've come in contact with, you'll have to tell me…but one of the other issues we were looking at was the impact of patenting of software, underlying elements of software applications. This may again be a question for one of the individuals actually working on preservation activities…

Sure. Yes, it probably is. If you drop me an email with a paragraph or so with the specific kind of issues you want to pursue and I'll bounce it around to see who the better person would be. That would be the most sensible thing.

That would be wonderful. Then maybe…I guess, maybe like looking at a more general level, what are the challenges that you see just from the viewpoint of your experience with your library…what are the major challenges that you see for digital preservation community with respect to digital rights management policy?

That's a very good question. It's a quite larger…

Yes, I know…there's several…just the largest, perhaps, that you see…the most significant challenges.

The thing that's…you may or may not know that actually with on the cusp of non-print legal deposit legislation in the [name of specific country removed] that comes into force in a couple of weeks time. And I wonder whether in some ways a legal deposit changes the rights issue entirely and presents all sorts of new and interesting questions of…but we actually don't know very much about yet. So that I suspect that actually if as I think you might anticipate more and more countries went to the direction of the legal deposit provision than one would expect those issues to start reproducing themselves in different places. And then we start together to get a better sense of what the issues in the implementation of legal deposit , what issues are for real. And it's hard to know what they are because we haven't starting doing it yet.

I was just about to ask you, you know, what do you think would be some of the first steps if that legislation went through that web archiving institutions would take in [name of specific country removed]…

I mean it's really ourselves in conjunction with the other legal deposit libraries. There are six legal deposit libraries. But they currently have dispensation for prints. And it's all being rolled over into non-print. There's a bargain struck in most legal formulations for legal deposit which restricts the access that we can provide to the material in return for a statutory overriding of the intellectual property issues…and the legislation gives us a good deal of indemnity against defamation and personal data and other things…because the state is asking us to do this in order for that to work it must stand in, put into law and understanding indemnity for those institutions. And so it simply remains: I don't really know. We don't…don't quote me on thisbut it remains to be seen how…because those provisions haven't yet been tested or repealed against in any sense or ways. It's very hard to know.

Right.

But that's where issues will come, I think. Our usual explicit permission, a way of doing things we have done so far, in a sense doesn't seem to throw up any rights issues. But actually this next step may do. It just remains to be seen how that works.

Do you think that there are…I mean, obviously, then best practices would also be modified in this way…do you think, memory institutions will engage in digital preservation or curation for the public good? What do you think are the best practices that exist currently to work around these legal issues?

If they can be called the best practices…my impression is that many institutions just tend to be very risk averse. And so they tend to do their risk assessment in relatively conservative way. So that actually if there is any risk at all then – even a small risk – then they tend to proceed in a more restrictive way….and so trying to forestall the issues at a policy level to begin with rather than on an item by item basis. I guess, you have a view on the German situation. That would be my impression of the way that most statutory memory institutions work because they are answerable to the state and they are part of the government so they need to be seen through white on white in this regard.

What do you think about the strategies of like e.g. the internet archive which is not risk averse at all…essentially.

What I think the IA can, as it were, because they are not an arm of the state. They fire first and ask questions later and see what happens. And actually, you know it's very helpful in many ways from an international web archive point of view. From a scholarly point of view, it's very good because that's what they've done. Because a whole lot of stuff got preserved that would have been lost otherwise…if they had been more concerned about the rights. So as a scholar, one would say that was a very good thing. But it's not…I can't imagine any national library adopting the same approach. Not even nationally, let alone internationally.

Simply because the risks of actually backing institutions are too great?

I think so. I think organizations like the British Library or the Bibliothéque Nationale or all the other national libraries they have issues in trust that they need to preserve. And actually to have been shown to have been knowingly not observed some sort of statutory or legal rights or what else, would be a major problem. Whereas the archive, it only does that. Its reputation is based on having done it already.

When you're speaking about reputation are you referring to the reputation among content providers or among…let me think…are you referring to the reputation that they have in terms of their, let's say, professionalism as an institution as such or with content providers that they can trust these institutions to be good stewards of their material and respectful of their intellectual property and…?

I think it's both. It's rather like…if an agency of governments which collected e.g. health information or information about criminal records or something…if it were somehow to expose, were not to observe all of the data protection legislation that was in place then it would be a major problem of trust in the integrity of that organization as part of government to have done that…to have been seen to do that. So that means institutions that are state funded directly have a very similar kind of reputation to manage. It's not really an issue about how they look after the content once they've got it in terms of print narrow preservation terms. It's more about a general scrupulousness about the law in all those aspects.

Right. Possibly this is kind of a shot in the dark but does your library consider all web pages as published material? Is everything considered published if it's not password protected or singled out by robots txt?

The legal deposit web station sees everything on the open web as in scope. Yes.

Ok. This was just a question. Ok. And one of the other possibilities that we considered was a sort of community driven rights management policy…we're not sure…we're trying to think of, you know, in terms of added value to the project…what might be other viable options to the ones that are currently existing. So let's say for example in the same way that Wikipedia's knowledge management is driven by public knowledge that we could harness a community of

users that would be content retrievers and content providers and any other sorts of interested bodies to help identify certain elements of rights management as part of our structure. Do you think that type of thing is a viable idea for a memory institution?

I'm not sure what exactly you mean.

Let's say for example if rights content is unknown…if we go for the example that a depositer will be responsible for saying "Yes, I have the rights to all the content contained on my blog and I give you permission to archive it." It would be possible that somehow perhaps this depositer missed something, you know, YouTube content or something like that…that you could rely on a community to help to identify those potentially problematic areas and to work with you…we are trying to think of ways that you could engage a community in being interested in web preservation and all of its aspects. So we are trying to identify this element in every part of the project. And we were wondering if there could be…if that seemed like any sort of realistic option for rights management policy.

Well, I think any kind of mechanism to gather experience of other people doing the same thing is in general a good thing. In that particular case, I tend to think that it would be an extremely risk averse institution that didn't accept the expressed permission of a site owner if they had specifically said and provided for that they were able to license all of their content and specify it. There is always a theoretical risk as you say…that they haven't understood it correctly or they just didn't know…I think if you can then cover yourself with something, you could then reversibly put together a notice - atake down policy of some sort. That ought probably to mitigate that risk. It would take…if got the drafting of the agreement right then I suspect that in that particular case…you would have to be very very risk averse not to accept that.

Perhaps, that's a better example for in the case where we would have opt out recommendations…because what we're doing is we are trying to create a set of recommendations for people that would use this platform. And ultimately it will be the end user who will decide what they actually want to do so we need to consider a lot of various possibilities. One would be that they are not seeking expressed permission before archiving – similar to the internet archive. And perhaps, maybe that community aspect could be one thing about that. But yes…sorry, I am just trying to come back to my questions here…
I am just looking through…I think a lot of the questions that I have remaining actually have to do with web preservation content…so I am wondering…let me go through…
Yes, I think actually the remaining issues that I have would be better directed toward a colleague that has more to do with the actual preservation activities involved in the web archiving. So if I send you an email I can even send you the questions that we would have for this individual and they could…

Yes. That makes sense. And I'll bounce it around to see who's available. I can't make any promises but I'll do what I can.

Sure. That would be wonderful.

Ok.

I really really appreciate you speaking with me today. Thank you for taking the time.

That's quite right. If you ‘re going to quote on me that explicitly you let me see it.

Of course.

Very good. Thank you.

Thank you very much.

No problem.

Interview Notes for "Sam Howard"

In general about web content crawling, according to your opinion. Do one need active acceptance or is passive acceptance normally sufficient before crawling?

In principle, active acceptance should be secured and potentially license agreements signed before anyone harvest and utilize web content in a formalized and commercialized way.
And of course, there are obviously rights of ownership involved in any published content on the net. However, from a practical and even legal point of view it's not possible to enter licenses discussions with all crawled material online. That would make the whole web & crawling concept collapse.

Robot.txt has for long been regarded as a basic signal of acceptance or denial to crawl a web site. Seeing beyond this principle, what alternatives to Robot.txt would you regard as guidance to scaled crawlers and search engines as of what to crawl or not?

Robot.txt is a basic principle, but doesn't include the concept of licensing. Within news crawling we have national initiatives that include licensing towards all or most publishers. Such is NLA in UK and Klareringstjenesten in Norway that helps simplifying the aspect of DRM in crawling – handling the terms of crawling of hundreds of news site.

Similarly one could hope for a set of predefined licenses that each content owner could choose from to publish and communicate license agreement for anyone interested in crawling their content.

Looking at blogs – wouldn't you say that publishing your content to a pingserver by act is in fact an acceptance to crawlers?

Yes, it's certainly the intention of connecting to a pingserver. If a blog author connects to a pingserver, it's by definition to ensure updates about your blog site are being captured by crawlers. So by action it's allowing crawlers.

In legal trials about crawling – eg Meltwater case in both US, UK and Norway – the issue of "fair use" has been raised. What is fair use in crawling?

One could say that crawling is a basic part of internet – needed to secure navigation in this vast amount of information. If crawling in general was illegal – it would have vast impact upon the entire internet.

But in fact, even In legal discussions, fair use have allowed crawling and indexing as such. The questions have been about how much to be represented when presenting the findings. Links has been acceptable, but length of abstract and even length of hyperlinked-headline has been said to be shortened when defining fair use towards news.

Fair use has also included the right to archive content for archiving for historical understanding and preservation in general.

How do you look upon consequences to society by limiting the right to license?

As mentioning it will definitely have negative consequences if online content where not allowed to be crawled and preserved. It will impact our ability to understand history. Also it will impact the

importance and influence of social media. As seen in Turkey these last week – publishing and getting spread out Twitter postings about riots and conflicts with authority influence the development of the society itself.

So what makes DRM so complicated? Wouldn't it be possible to have 1-5 alternative "Term of use" or license agreement which could cover the entire right of crawling and utilizing blog content?

Yes, one would think that could be define by each blog author to regulate further use of the content.

And one could then claim that imposing such "term of use" is the responsibility to all authors in order to protect their content.

However, what's making DRM and legal aspect on internet so complicated is that there is so many layers of rights and connection to potentially different jurisdictions. This increases the risk of conflict of terms and rights – as well as interpretations.

For instance – while the blog author may relate to the law of his citizenship, he might also have to relate to the law of where the content is origin, or the origin of the content described. But also the blog might be hosted in another country with different laws. And the reader or crawler could be located in another jurisdiction. And the repository or search engine displaying the content might operate under yet another jurisdiction.

And the author might allow the content to be crawled and archived, while the blog platform itself has a non-crawling term-of-use- clause.

The obvious problem is then – can one ever be 100% sure there are no legal issues when crawling and defining a DRM?

No, I'm afraid the lack of standard international laws within content handling online is making guarantees of a legal risk free crawling impossible.

But I find the discussions you have about DRM within the Blogforever-project can become highly valuable.

I would be very interested to see what you are able establish within this Blogforever project.

Appendix C. Interview Questions for Section 6

These are questions that were used in the interviews presented and discussed in Section 6.

1. How frequently do rights holders bring legal action against (e.g. sue) public institutions in the UK or abroad which are preserving digital materials? And do those legal actions result in damages or merely a demand that the content be taken down which is satisfied by the content being taken down?

2. What kind of legal issues (related to rights, statutes, licenses, surveillance, privacy, data protection) have your library and other web archives experienced? How were these resolved? Did any of these issues lead to serious risks to the archive?

3. What are the implications of multi-jurisdiction rights on the preservation of content (e.g. what happens if a blog from the UK contains material that infringes rights in France, the US, and New Zealand)?

4. How do web archives determine web information under their jurisdiction? Is it solely according to domain name in the URL? For example, if a person in the UK uses a public blogging platform (such as Wordpress) to post material on the web, to which jurisdiction does this belong?

5. What would be your recommended strategy for developing a rights management policy for an archive/library/repository aiming to preserve the content of blogs?

6. Have there been notable problems and/or risks specific to the "opt-in until opt-out" principle applied by the Internet Archive?

7. How is contract law, in the form of software and content licensing, reshaping the digital rights landscape? What impact will this have on preservation and curation of digital content by public organisations?

8. As you are perhaps aware Blogforever is focused on developing tools that will support the preservation of blogs. Blogs often contain an interplay of digital materials, some content and some software, and this interplay must be maintained in order to ensure the authentic preservation of the blog. What best practices exist to enable preservation organisations to maintain and provide access to web content such as blogs?

9. Are there any legal cases in the last five years in the UK or abroad which might be especially illuminating in helping us to understand issues related to digital preservation and the law?

10. What is the impact of patenting of software or underlying elements on which software applications depend (e.g., algothrims [say Amazon's One Clip patent], interface design patents) on preservation? What are the best practices for mitigating any impacts?

11. If a memory institution were to accession and preserve a blog which unknown to that institution had already infringed intellectual property rights (e.g. copyright, patent) of others, can the memory institution have a realistic expectation of absolving itself of legal (whether criminal or civil) responsibility for the infraction by claiming it had acted in good faith in assuming that the blog creator had addressed all the rights issues?

12. Can blogs that use a variety of materials sourced from different content providers (e.g. a blog that discusses a film and uses clips from a number of other films or audio recordings to support its arguments) ever be considered to be an orphan work? (The problem here is that it includes other materials which have rights issues associated with them--they belong to others)? If yes, under what circumstances?

13. How does the EU Directive on Databases impact on the preservation of blogs as blogs are normally represented as outputs of commercially held databases?

14. It would appear digital objects can be covered by multiple rights simultaneously (e.g. patent, copyright) and fall under multiple legal domains (e.g. civil and criminal laws). How does this layering of rights affect preservation practices?

15. What changes do you see coming in the future with respect to rights management in the context of web information?

16. Where memory institutions are engaged in digital preservation/curation for the public good (e.g. memory of society), are there any best practices to work around the legal issues?

17. What are top three challenges you see for the digital preservation community with respect to digital rights management policy?

18. What are the top three challenges you see for web archiving with respect to digital rights management policy?

19. What are the top three challenges you see for blog preservation with respect to digital rights management policy?

20. Does your library and other web archiving institutions across Europe consider all webpages as "published" material (for example, is everything considered "published" if it is not password protected or singled out by robots.txt; assuming these are excluded)?

21. In your opinion, is a community driven rights management policy (as opposed to one driven by an institution or public body) a viable option (for example, in the same way that Wikipedia's knowledge management is driven by public knowledge)?