SEVENTH FRAMEWORK PROGRAMME
FP7-ICT-2009-6

BlogForever
Grant agreement no.: 269963

BlogForever: D3.1 Preservation Strategy Report

Editor:	**Yunhyong Kim, Seamus Ross**
Revision:	First Version
Dissemination Level:	Public
Author(s):	Yunhyong Kim, Seamus Ross, Karen Stepanyan, Ed Pinsent, Patricia Sleeman, Silvia Arango-Docio, Vangelis Banos, Ilias Trochidis, Jaime Garcia Llopis, Hendrik Kalb
Due date of deliverable:	30 September 2012
Actual submission date:	30 September 2012
Start date of project:	01 March 2011
Duration:	30 months
Lead Beneficiary name:	University of Glasgow (UG)

Abstract:

This report describes preservation planning approaches and strategies recommended by the BlogForever project as a core component of a weblog repository design. More specifically, we start by discussing why we would want to preserve weblogs in the first place and what it is exactly that we are trying to preserve. We further present a review of past and present work and highlight why current practices in web archiving do not address the needs of weblog preservation adequately. We make three distinctive contributions in this volume: a) we propose transferable practical workflows for applying a combination of established metadata and repository standards in developing a weblog repository, b) we provide an automated approach to identifying significant properties of weblog content that uses the notion of communities and how this affects previous strategies, c) we propose a sustainability plan that draws upon community knowledge through innovative repository design.

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	21/08/2012	Drafting process (Version from incorporating contributions throughout WP3 Task 3.1 by the people listed in a author's list on the cover page).	Yunhyong Kim
0.91	27/08/2012	Drafting chapters 5 and 6	Yunhyong Kim
0.95	03/09/2012	Last stages of drafting	Yunhyong Kim
0.99	26/09/2012	Draft for review by WP3 and the project management.	Yunhyong Kim
1.0	30/09/2012	First version of the deliverable	Yunhyong Kim
1.1	25/09/2013	Updated version of the deliverable	Yunhyong Kim

Table of Contents

Executive Summary

This report describes the strategies developed within the BlogForever project to support the preservation of weblogs. It describes the work carried out to meet the objectives detailed in the BlogForever Description of Work (DoW) as being part of Task 3.1 Development of the Preservation Strategy. According to the Project Description of Work[1], the main objective of this task is to develop a preservation strategy that will

1. "ensure the reliable maintenance of weblog data, and, further, the long-term accessibility of digital content objects deemed to have enduring value";
2. "assess the risks for loss of weblog content and related digital content objects";
3. "determine preservation actions";
4. "determine the appropriate metadata needed for each object type, and ensure access to this content";
5. "review existing digital preservation strategies to select the most appropriate one for weblog digital preservation";
6. "ensure that the preservation method selected will also retain item inter-relations";
7. "guarantee the successful fostering of the preservation strategy by the final repository system".

The advent of weblogs can be placed at a turning point in the development of the Internet, when the Internet changed from the transmission medium it used to be to the "social" communication medium it has become, a transformation made possible through the proliferation of interactive channels such as Twitter, networking channels such as Facebook and LinkedIn, and, channels of sharing personally curated information such as Pinterest.

While there have already been many web archiving initiatives with varying themes and focus, few of them provide adequate support for the "social" dimension of what we have come to call the *social web*. It is well recognised that, before we can identify web "objects deemed to have enduring value", assess "risks of loss" associated with these objects, and determine "preservation actions" as a strategy, we must understand what range of objects are in a target collection. As a result, recent projects have focused on the range of mime types, required characterisation processes and associated problems of scalability (with respect to heterogeneity and volume of the collection). To capture the social dimension of weblogs, however, it is paramount that we understand how each component type is used within the weblogs, how they are inter-related to each other (both spatially and temporally), and, which of these form the expected defining features of the community that produced the blogs.

Existing and new preservation strategies must be selected to support the preservation of these features with full awareness of their impact on complexity and scalability of preservation processes. The emphasis of developing a strategy cognizant of such pragmatic features is intended to be based on the observation that these features often provide the context of creation central to supporting the appraisal of blogs with respect to authenticity, integrity, reliability, and completeness of information.

Most previous digital preservation strategies have been reliant on knowledge and software engineering solutions (for example, development of standards, significant properties generated by knowledge experts, and defining best practice workflows). These engineering approaches can be rigid to be suitable only for a selected context (i.e. does not adapt easily to changes), expensive to modify for changing situations and communities, and could be subject to preservation actions themselves over time (that is, a preservation plan for the preservation system itself might become necessary – potentially leading to infinite recursion). This poses questions of their true long-term

[1] Page 10, Part A, BlogForever Project Description of Work.

sustainability. These approaches also place the burden of feasibility and practicality, on the design of the technical infrastructure. For example, solutions for handling scalability, availability of resources, simplicity of implementation, and ease of maintenance tend to rely mostly on distributed methods such as that represented by cloud computing (cf. approaches being proposed by the SCAPE project[2]).

However, discussions on specialist blogs[3] show that, as tasks become more complex the demand on communication and coordination required between distributed processes can create a bottle neck in processing speed as much as the sheer volume and/or heterogeneity of data. Preservation processes in the web environment are both big and complex: the large volume of web information naturally suggests scalability issues, but, on top of this, the identification of subcomponent object types and independent preservation actions that might be required for each object type (e.g. characterisation, migration, link update) portends that it is not even clear that all the processes can be handled within the time frame that the rapidly changing and growing web demands.

This is why recent projects (such as SCAPE, ARCOMEM[4] and LAWA[5]) have opted to consider distributed computing methods as a solution. Here, we offer the following observations:

- Distributed computing often involves a partition of data which is independent of the syntax, semantics, and pragmatics used by the information creator. While the reconstruction of the information may be currently successful, this, nevertheless, presents a risk to integrity and completeness over the long-term.
- Developing approaches to analysing focused community practices, that are transferable across communities, can support solutions to issues of feasibility and practicality, while minimising requirements for distributed storage.
- Any digital preservation strategy for maintaining community-driven social network media should support technologies, significant properties, and user requirements resulting from analyzing focused designated communities.

In light of these observations, we propose to meet the demand of increasing volume and management complexity by first profiling clusters of weblogs with respect to focused weblogging communities characterised by the technologies they share, the social network they form, and how they organise their information. This approach will immediately reduce both volume of data and the scope of technologies that need to be supported for the target community. Depending on the number of community clusters within the scope of the archiving organisation, this can also be combined with distributed and parallel computing methods such as cloud computing, for added speed.

The central concept here is that we are proposing an approach to detecting and profiling the features of online communities that is transferable across a variety of communities. We are not proposing the development of independent solutions for each community; we are suggesting a general approach that can be used to tap into arbitrary communities on the web. This aspect distinguishes it from initiatives that provide solutions to problems specific to a selected community (cf. the software development mashups organised by SPRUCE project[6]).

Understanding communities through the analysis of network structures and/or topic categories and tags is not new (cf. approaches to content selection used by the ARCOMEM project). However, this kind of analysis is rarely combined with the analysis of technical conventions within a community. Here we make a step forward by bringing an analysis of the technical conventions to the table.

[2] http://www.scape-project.eu/
[3] http://highscalability.com/blog/2010/3/30/running-large-graph-algorithms-evaluation-of-current-state-o.html
[4] http://www.arcomem.eu/
[5] http://www.lawa-project.eu/
[6] http://wiki.opf-labs.org/display/SPR/Home

It intuitively makes sense that members of the same community might use similar file formats, browsers, blogging platforms, technologies, and tools. Examining these conventions not only helps in developing digital preservation strategies that support selected designated communities, but, also serves as a counter balance against value assigned to information on the basis of general popularity (see for example the algorithm that Technorati[7] uses to assign "authority" to blogs) by highlighting information produced in an uncommon file format which is, nevertheless, prominent within a selected community.

Another advantage of a weblog repository that is cognizant of online communities is the potential that the repository has to attract contribution from the target expert knowledge community: this could be in the form of providing general feedback, refining metadata, alerting the community to access problems, requesting missing information, and contributing solutions to problems. The next stage of digital preservation research must include a programme of developing ways of redirecting expert attention. We advocate such a programme because of the unavoidable recognition that problems of complexity and scalability with respect to preservation processes will only increase and automation and distributed computing will not be sufficient as a solution to the problem.

In summary, the report culminates in three contributions to the current research landscape:

- automated approaches to identifying significant properties of weblog content that draws upon the notion of communities,
- transferable practical workflows for applying a combination of established metadata and repository standards in developing a weblog repository, in light of the identified properties, and,
- proposals for a sustainability plan that draws upon community knowledge through innovative repository design.

The report is structured into seven chapters: a discussion of why we would want to preserve blogs, the preservation activity that BlogForever supports, and the contributions of the work presented in this report (Chapter 1); a review of previous strategies and projects relevant to this report and why they are not adequate for blog preservation (Chapter 2); a more detailed description of what it is we are aiming to preserve in BlogForever (Chapter 3); automated profiling of weblogging communities and how this helps weblog preservation (Chapter 4), a survey of standards explored and/or adopted within the BlogForever repository (Chapter 5); a discussion of BlogForever repository in relation to repository standards (Chapter 6), culminating in a proposal for weblog preservation strategy (Chapter 7) and suggestions for future work (Chapter 8).

[7] http://www.technorati.com

1 Introduction

1.1 Why Preserve Weblogs?

Information on the Internet can be rather fleeting: page links that exist today can easily disappear tomorrow and resources embedded within webpages become misplaced resulting in the loss of information integrity and completeness. In some collection-based studies, it is shown that up to 38% of page links can disappear from the live web over the course of four to five years[8]. It has been pointed out that "link rot" is often a consequence of human incompetence rather than a result of failing digital technology (Berners-Lee 1998). Nevertheless this observation has not led to a noticeable decrease in the volume of lost links. As a response to the lack of persistence within the web, many archives have taken on the mission to harvest and store a selection of webpages deemed relevant to their governing body, organisation, and/or company (see the list in Chapter 2).

One of the first to store periodic snapshots of the web over time has been the Internet Archive[9], founded by Brewster Khale in 1996. There is, however, a noticeable loss of resources embedded within the pages stored at the Internet Archive, which increased as more external media made its way into the World Wide Web. Some recent initiative and projects have tried to provide solutions for the problem of resources that go missing by providing resources that closely match the requested resources with respect to the time of publication[10]. While the solution seems promising, there are two immediate observations: the methodology depends on linking the resource to a time gate and it can only return the closest match. While the time gate concept provides a pointer to a central URI to collect versions of a resource, it is not clear that this cannot go astray. It is also not clear how good the match will be in the long term when most of the versions might be lost and only versions at wide apart intervals are available. It also needs to be verified whether the solution is sustainable as forms of *web communication* become increasingly complex.

The ephemeral and fragmented nature of Internet resources as publicly available information should not be the sole driving force for the need for web archiving and preservation. It is, in fact, often, essential that society is allowed to forget[11]. For example, a given piece of text could just as easily mislead as much as inform society. The decision to archive and preserve web information must be driven by information value, not as we perceive it in the present, but as a potential source to trace accountability, to revisit history and culture, to discover new knowledge, and to improve the quality of life.

The potential of social network data as a basis of social policy development, as a potential record of accountability, sometimes shaping social behaviour and effecting social change is increasingly being recognised[12]. Likewise, the informational value of social network data to support new discovery is rearing its head: for example, it could be re-purposed to provide the essential life style information that would bring the information required to improve medicine[13]. As a social and literary phenomenon, weblogs (or blogs[14]) have been of interest almost since their inception: there have been analyses of the weblog in genre studies (for example, Herring et al. 2004; Miller & Shepherd 2004), in social network and cybernetic culture studies (for instance, Caverlee & Webb 2008, Wilkinson & Thelwall 2010), and. more recently, in exploring the relationship between blogging and self identity (e.g. Siles 2012). Some have mentioned weblogs as valuable records of historical

[8] http://www.theverge.com/2012/5/15/3021913/chesapeake-digital-preservation-group-link-rot-report
[9] http://archive.org
[10] http://mementoweb.org/
[11] http://greenmediabox.eu/archive/2012/06/28/data-protection/
[12] http://www.technologyreview.com/featured-story/428150/what-facebook-knows/
[13] http://www.ihealthbeat.org/perspectives/2011/the-rise-of-social-media-and-participatory-medicine.aspx
[14] We will use the terms weblog and blog interchangeably throughout the report.

events for safe keeping (see Chen 2012) and their value as public records is rising as governments have been increasingly using these channels for communication with their constituents[15].

Some other merits of preserving weblogs have also been observed within other project deliverables of the BlogForever project (e.g.BlogForever:D6.3 Market Analysis). The potential value of having long-term access to weblogs seems to be clear. The crucial question is whether we can reasonably store them within a repository, providing *sufficient* evidence for evaluating the authenticity, integrity, reliability, and completeness of the given information. Even if we are able to successfully harvest all the necessary elements of a page into a weblog repository and ensure their continued access, the links within the pages will be subject to the link rot phenomenon we discussed earlier. This brings up the question of where the responsibility of the harvesting organisation lies. Pages associated to valid links and HTML markup, still need to be examined for integrity (screening to detect corruption during transmission and to manage changes or modifications), furthermore, the availability and integrity of embedded objects will also need to be investigated to estimate suspected information loss (see Bar-Yossef 2004, for a study of the decay of links and resources found within randomly selected webpages).

Assuming all of this is doable, it is still questionable whether or not the preservation objectives to enable the continuation of semantic and pragmatic integrity have been met. How would we measure the information loss caused by the information gap introduced by the loss of unharvested links, third party content and inter-related components? Further, will it be possible to provide the necessary evidence to measure reliability and authenticity? And, finally, can there be a digital preservation strategy to achieve these aims that are scalable to the volume of information now being produced online? These are the questions we want to attempt to answer within this report.

To achieve this, we have tried to elaborate on the notion of blog communities. A distinguishing feature of blogs (in comparison to other webpages), is a keen sense of an underlying *community of practice* (Lave & Wenger 1990; Hanson-Smith 2012) that emerges through the many subject specific blog portals that are now visible online (e.g. for social sciences, SocioSite[16]; for physical sciences, ScienceSeeker[17]; for mathematics, Mathblogging.org[18]; and, for fashion, Independent Fashion Bloggers[19]; as well as, sites for searching a broad spectrum of blogs, such as Technorati[20]). As one discussion on the mathematics blog, N-category Café[21], indicates, the medium has potential to provide insight into the history and philosophy of science, by making explicit the scientific processes as they happen within the community. This trickles down to all levels of the community to improve conversations with students, provide a meritocratic platform for open discussion, and provide a platform for "publications before publication". In fact, cutting edge mathematical problems have been solved in collaboration online, eventually resulting in a formal publication[22]. Weblogs have also been mentioned as a medium for redirecting expert attention to immediate problems, bringing faster progress and advances to areas where the need is greatest.

To develop a strategy for such community processes we need not only to take steps to provide access to the individual components of the weblog but also the connections that exist between weblog content. It is difficult to imagine that a system that retrieves target items on the submission of keywords or metadata would be sufficient to allow future demands in the analyses of weblogs. This is not merely a matter of providing end-user analyses tools, as suggested by the International

[15] http://www.records.ncdcr.gov/guides/bestpractices_socialmedia_local_2010412.pdf
[16] http://www.sociosite.net
[17] http://scienceseeker.org/
[18] http://www.mathblogging.org/
[19] http://heartifb.com/
[20] http://technorati.com/
[21] http://golem.ph.utexas.edu/category/
[22] See polymath project blog description at Wikipedia: http://en.wikipedia.org/wiki/Polymath_Project

Internet Preservation Consortium[23], but preserving the correct scope of weblog features in the first place, to reflect the community process, network, and information sharing activity, and providing them in a way that is open to such analyses tools (for example, a lot of these tools require data in plain text). When we observe the changes that have taken place with respect to how we access information over the last twenty years it is clear that we have moved from the practice of searching a catalogue for explicitly referenced material to search based on several levels of stratification and connections. The strategy proposed within this report aims to suggest a workflow for extracting significant properties of a weblog that focus on maximising the probability of recovering the stratification and connections associated to target communities.

1.2 BlogForever Objectives

The current framework of the BlogForever project is focussed primarily on capturing, storing and rendering the *content* of a weblog, rather than its look and feel or behaviour. We anticipate the content to be the material published by bloggers, largely in the form of posts, comments, and associated metadata. The BlogForever service will deliver this textual content through the access mechanisms of the repository. Secondary to the text content will be the media content, in the form of images and attachments.

This framework means that the more complex behaviours of blogs (such as external linking, content that relies on an external database, embedded content, GIS data, and further complex objects such as 3-D images) are out of scope of the current iteration of the capture and rendering strategy. However, this does not invalidate D3.1's evaluation of blogs as complex objects, and this evaluation will be needed for future iterations of the BlogForever service.

While the digital preservation of the above mentioned complex objects are a challenge in themselves, it is our view that the immediate challenges that surface in the weblog context are in:

1. Establishing a feasible management strategy for the volume and variety of different types of objects that appear within the weblog pages, and,
2. Sustaining the integrity of inter-relationships between these objects and the weblog page.

By examining the complexity and scalability involved in the relationship with respect to webpage text, images, audio, video and documents, we hope to shed light on further refinements involving increasingly complex objects.

Our shared understanding of a successfully preserved blog is an Archival Information Package that contains:

- An object, or set of objects, that when assembled correctly through a suitable platform will provide a rendition of the blog text and media content as captured by the spider
- Sufficient descriptive metadata that describe this content, including its original provenance, location and date of capture
- Sufficient technical metadata that identify, measure and declare the significant properties of each digital object in the package
- Records of any preservation and curation actions carried out on these objects, to be retained as preservation metadata

Our understanding is that the BlogForever service, within the scope of this project, will not constitute a "permanent collection" of blogs. Rather, the end result is more likely to be a *demonstrator service*; to prove that it is possible to capture and render blog content and put it into a preservable state within the context of a preservation-friendly repository system. Digital

[23] http://netpreserve.org/resources/web-archives-futures

preservation is not simply about technology, but also requires organisational or institutional support, and the provision of necessary resources. [24]

According to the Project Description of Work[25], the main objective of Task 3.1 of the BlogForever project is to develop a preservation strategy that will

1. "ensure the reliable maintenance of weblog data, and, further, the long-term accessibility of digital content objects deemed to have enduring value";
2. "assess the risks for loss of weblog content and related digital content objects";
3. "determine preservation actions";
4. "determine the appropriate metadata needed for each object type, and ensure access to this content";
5. "review existing digital preservation strategies to select the most appropriate one for weblog digital preservation";
6. "ensure that the preservation method selected will also retain item inter-relations";
7. "guarantee the successful fostering of the preservation strategy by the final repository system".

To answer 1, it was important to define content deemed to be of enduring value. Enduring value is an elusive notion that is in constant flux, however, there seemed to be five main factors that was deemed viable as a mechanism for capturing value: technical information that would allow future users (machines and/or humans) to access the information, information that traces the historical development of activities within a weblog community, aspects that conform to the core object structure of a weblog, aspects that meet the needs of immediately foreseeable stakeholders of the information.

1.3 Contributions of This Report

At any fixed point in time, the weblog page is not much different from any other webpage. The "digital content object deemed to have enduring value" that distinguishes weblog pages from other webpages is in the social interaction that it generates. The social aspect is not so much characterised by any fixed aspect, such as the variety of different digital object types found within the blogs, but, by the inter-relationships between pages and other pages, and their subcomponents, how they are used, and how these change over time. The risk with respect to losing weblog content is in the loss of these connections and use contexts, as much as the isolated information objects that are contained within the weblogs.

The current report culminates in three main contributions to the current research landscape:

- an automated approach to identifying significant properties of weblog content that uses the notion of communities,
- transferable practical workflows for applying a combination of established metadata and repository standards in developing a weblog repository, and,
- a sustainability plan that draws upon community knowledge through innovative repository design.

In line with workflows that have been developed within the DELOS network of excellence[26], and refined within projects such as PLANETS[27] to test preservation strategies, the approach taken in

[24] See the Three-Legged Stool model devised by Nancy M. McGovern and Professor Anne R. Kenney, which consists of organizational infrastructure (the "what"), technological infrastructure (the "how") and a resources framework (the "how much") of building an organization's digital preservation program.
[25] Page 10, Part A, BlogForever Project Description of Work.
[26] http://eprints.erpanet.org/48/
[27] http://www.planets-project.eu/

this report in developing a preservation strategy is that, in order to test preservation strategy alternatives, we first have to characterise the object we are trying to preserve, that is define the *significant properties* of weblogs. The PLANETS workflow was subsequently supplemented with approaches to user validation and user requirement analysis by the UK Digital Curation Centre (DCC 2008) and the InSPECT project[28]. There has also been further work in the direction of object characterisation within the XCL project[29].

Here we combine a macro level analysis of significant properties on the level of the data model (Section 3.3) with a micro level analysis (Section 3.4), by a study of how the micro level properties manifest in the macro level structures with respect to weblogging communities (Chapter 4).

In summary, we adopt three lines of investigation in parallel to define the *significant properties* of weblogs as a first step towards preservation strategy testing:

1. We map the BlogForever user requirements analysis (Kalb et al. 2011) to the data model to determine significant inter-relations between components of the weblog on the macro level (Section 3.3).
2. We draw upon the BlogForever weblog survey to identify the most prominent object types found within weblogs (Section 3.1) and we apply previous studies of the identified digital object types to determine significant "technical" properties of these digital objects at the micro level (Section 3.4).
3. We bridge the above two approaches by presenting an analysis of features that characterise the community which highlight how the micro level digital objects are used within the context of the macro level structure of the weblog (Chapter 4).

This distinguishes BlogForever from other research efforts where only one of these approaches has been attempted. In addition to this we present the recommended practice for metadata assignment and encoding standards (Chapter 5) that we propose for independent digital object types that have been found to be most prolific within the currently available weblogs. We further, present a preservation services requirement workflow (Chapter 6). In recognition of the widely accepted reference model for an Open Archival Information System (OAIS), this workflow serves as a OAIS-like repository workflow that is suitable for weblogs.

The analysis of weblog features presented in Chapter 4, also brings to light the question of whether the concept of "representative data" is sufficiently explored in the current practice of preservation testing, and whether, given the complexity of inter-connections between items (across space and time) in weblogs, preservation processes can be made to be scalable. As a solution to possible scalability risks, we suggest possible innovation in repository design that could redirect expert community knowledge back into the repository to create a community driven preservation strategy.

1.4 Structure of the Report

The remaining content of this report is divided into seven chapters. Chapter 2 presents a brief history of web archiving in the context of digital preservation and follows this with a survey of recent projects and work that are relevant to the development of the preservation strategy reported here. Chapter 3 describes the target of preservation, namely, weblogs, and their "significant properties". We discuss the risk of information loss, in Chapter 4, focusing on problems that surface in relation weblog complexity and relate this to the notion of blogging communities. The metadata recommended for recording the properties of weblogs is presented in Chapter 5, followed by a discussion of the BlogForever repository workflow in the context of the OAIS (CCSD 2002) and

[28] http://www.significantproperties.org.uk/
[29] http://planetarium.hki.uni-koeln.de/planets_cms/about-xcl

DRAMBORA[30]. The integrated preservation strategy consisting of these components is summarised and augmented in Chapter 7. The report is concluded in Chapter 8 with a summary of our contributions, lessons learned and description of future work.

[30] http://www.repositoryaudit.eu/

2 Previous Work: Review and Criticism

In this section we have reviewed some of the projects and initiatives deemed most relevant to discussions arising within this report. A more comprehensive collection of preservation projects and web archiving initiatives have also been surveyed as part of BlogForever deliverable D6.3 Market Analysis. Here we mention some new initiatives that were not covered in the Market Analysis and focus on those results that are especially relevant to web archives and strategy development for the preservation of social network media. The two surveys are intended to complement each other.

2.1 A Brief Overview of Web Archiving in the Context of Digital Preservation

Observations regarding the preservation of digital material as a challenge can be traced back, at least, to the 1970's (e.g. Dollar 1971). It started picking up speed, however, a little over twenty years ago (see Table 2.1-1). Despite the feeling that web archiving is fairly new, an interest in preserving information from the web is almost as old as the interest in digital preservation itself. The internet archive was already founded in 1996, around the same time that the Research Library Group[31] (RLG) Task Force on Digital Archiving report was produced, and Margaret Hedstrom's well-recognised paper (Hedstrom 1997) on digital preservation was published.

Table 2.1-1 Historical overview comparison of development: web archive versus digital preservation

1990	1992	1994	1996	1998	2000
Besser (1990) Visual Images at UC Berkeley.	Kenney & Personus (1992) Cornell/Xerox Commission on preservation and access.	Research Libraries Group Task Force on Digital Archiving (1994) Rothenberg (1995) "Ensuring the Longevity of ..."	RLG Task Force (1996) "Preservation of Digital Information" Hedstrom (1997) "Digital Preservation: a time bomb for ..." Making of America II (1997) Internet Archive (1996) Bibliotheca Alexandrina (1996) UK National Archive (1997) Sweden (1997)	Ross & Gow (1999) "Digital Archaeology ..." Berners-Lee (1998) CAMIiLEON (1998) CEDARS (1999) LOCKSS (1999) Austailia Pandora Archive (1999) Newzealand (1999)	Ross (2000) "Changing Trains at Wigan ..." Granger (2000) "Emulation as a Digital Preservation Strategy" Digital Preservation Coalition (2001) ERPANET (2001) METS encoding (2001) Library of Congress Minerva (2000)
2002	**2004**	**2006**	**2008**	**2010**	**2012**
Thibodaeu (2002) "Overview of Technological Approaches .." Reference Model for an OAIS (2002) IIPC (2003) PREMIS working group (2003) France Bnf (2002) UCLA (2002) Japan (2002) California (2003)	Wilson (2005) "A Performance Model ..." Knight (2005) "SHERPA-DP OAIS-Report .." Lavoie (2004) "Thirteen ways ..." DCC (2004) PrestoSpace (2004) PRADIGM (2005) Internet Memory (2005) PREMIS data dictionary (2005) Canada (2004) UK British Lib (2005) Iceland (2005) Denmark (2005) Korea OASIS (2005) Catalonia (2005)	DINI-Certificate (2006) Dobratz (2006) "Catalogue of criteria for trusted ..." Ambacher (2007) "TRAC" DRAMBORA (2007) CASPAR (2006) Digital Preservation Europe (2006) PLANETS (2006) PROTAGE (2007) InSPECT (2007) InterPARES (2007) SHAMAN (2007) Linked Data (2007) Finland (2006) Netherlands (2007) Slovenia (2007)	Data Seal of Approval (2008) Dappert & Farquhar (2009) "Significance is in ..." nestor Catalogue (2009) LiWA (2008) Parse.insight (2008) Papyrus (2008) 3D Coform (2008) LAWA (2009) Dublin core as ISO standard (2009) KEEP (2009) Columbia U (2008) Switzerland (2008) Austria (2008) Harvard U (2009) Ina (2009)	PersID (2010) Open Planets Foundation (2010) Wf4Ever (2010) Memento (2010) BlogForever (2011) APARSEN (2011) ARCOMEM (2011) ENSURE (2011) SCAPE (2011) SCIDIP-ES (2011) TIMBUS (2011) U of Michigan (2008)	SPRUCE (2012)

This was soon followed by the CAMiLEON[32] (Creative Archiving at Michigan and Leeds Emulating the Old On the New) and CEDARS[33] (CURL Exemplars in Digital ARchiveS) projects.

[31] http://en.wikipedia.org/wiki/Research_Libraries_Group, now merged with the OCLC (http://www.oclc.org/default.htm).

[32] http://www2.si.umich.edu/CAMILEON

The CAMiLEON project, for example, was one of the first to compare emulation and migration and the impact of each on the digital repository. Around this time, the UK National Archive[34] started collecting and archiving UK government websites and Australia's web archive Pandora[35] was established (in operation since 1999) along with New Zealand's Web Archive[36].

The ephemeral nature of web resources has encouraged a number of other actions taken (e.g. Internet Archive[37], HTTP Archive[38] or International Internet Preservation Consortium[39]) for ensuring their long-term accessibility and preservation. The main goal of these initiatives is to prevent the loss of information and knowledge available on the Internet and make it accessible for users and generations to come. By archiving web resources, these initiatives aspire to offer access to information even after it disappeared from the Web.

The rationale for web archiving, however, is not limited to preservation of individual resources over time. The Web, which constitutes a platform for debate, creation, collaboration and social interaction, reflects many aspects of our society. It becomes a historical and cultural necessity for larger archiving initiatives to capture these characteristics of the Web. Records of collective heritage encompassed in the Web, rather than individual resources, should be of interest to archivists (Masanès, 2006). Preserving this heritage will not only provide access to historical artefacts, but also record the evolution of the medium – the Internet of the past and the present (Brügger, 2011).

Web archiving initiatives, such as ARCOMEM[40] or LiWA[41], have been increasingly trying to create solutions for social media archival situations. However, current preservation initiatives do not make adaptive provisions for dynamic and interactive environments such as blogs and social networking media. Instead, they tend to focus on various levels of version control and neglect deeper interactive aspects coming from networks, events and trends.

Table 2.1-1clearly shows a growing interest in web archiving, and, in recent years, in social web archiving, evidenced by a growing number of national web archives (the most notable increase is observable in 2004) and projects funded in the area of web archiving, web analytics, web resource management (for example, LiWA[42], LAWA[43], BlogForever[44], ARCOMEM[45], and Memento[46]). A list of the projects and the many web archives and their URI can be found in Appendix G.

Table 2.1-2 Digital Preservation Research Identified by the DigitalPreservationEurope Research Road Map.

[33] http://www.ukoln.ac.uk/services/elib/projects/cedars/
[34] http://www.nationalarchives.gov.uk/
[35] http://pandora.nla.gov.au/
[36] http://www.natlib.govt.nz/collections/a-z-of-all-collections/nz-web-archive
[37] http://www.archive.org/
[38] http://httparchive.org/
[39] http://www.netpreserve.org
[40] http://www.arcomem.eu/
[41] http://liwa-project.eu/
[42] http://liwa-project.eu/
[43] http://www.lawa-project.eu/
[44] http://blogforever.eu
[45] http://www.arcomem.eu/
[46] http://mementoweb.org/

Table 1: Simplified Crosswalk Analysis Matrix

	UEI	PDI	DPNU	SoDP	IAT	I2S	eScience	Cyber	DigiCult	Erpanet	Warwick	DRR	OSI[3]
	1991	1996	1998	2002	2003	2003	2003	2003	2004	2001-2004	2005	2006	2007
Digital Object Level													
Migration		+		+	++	+							
Emulation				++	+						+		
Experimentation		+				+							
Registries and repositories					+++	++++				+	++		+
Complex Objects	+			+	++	+		+	++	++	++		
Significant properties			+	+	++	+	+						
Authenticity				++	+++++					+	+		
Acceptable loss					+			+					
Collection Level													
Interoperability				+	+++	+					+		
Metadata		+			++++								
Management					+++	+							
Standardisation			+		++		+	+		+	++	+	
Media Types					+		+		+				
Repository Level													
Tools and architectures		+			+		++						+
Benchmarks		+			++++		+						
Hardware Issues								++++			++		
Storage		+		+	+	+++		+		+	++		
Trust		++	++	+++	+	+++			++	++++			
Scalability					+++			++					
Sustainability				+	+++					+	++++	+	
Planning		+		+			++			+	+	+	
Repository Management							+			++	+		
Cost					+		+			+	+		
Process Level													
Access				+		+		+	++		++		
Automation				++	+++++	++++	+		+++++		+++++		
Monitoring				+									
Organisational Environment													
Creation and use	++	+					+						+
Legal Issues		+	+	+++			+	+		+			+++
Collaboration		+++		+++	+			++					++

Table 2.1-3 Keys to Research Initiatives Listed in Table 2.1-2.

- **UEI** – *Understanding Electronic Incunabula: A Framework for Research Electronic Records* [9] by Margret Hedstrom, 1991.
- **PDI** – *Preserving Digital Information* [17], edited by John Garrett and Dona Waters, 1996.
- **DPNU** – *An Investigation into the Digital Preservation Needs of Universities and Research Funders* [15] by Denise Lievesley and Simon Jones, 1998.
- **SoDP** – *The State of Digital Preservation - An International Perspective* [5] contains articles by various authors, 2002.
- **IAT** – *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation* [3] was published by the NSF in 2003.
- **I2S** – *Invest to Save* [4] was prepared for the NSF-DELOS working group on digital archiving and preservation in 2003.
- **eScience** – *e-Science Curation Report* [16] by Philip Lord and Alison Mcdonald was published in 2003.
- **Cyber** – *Revolutionizing Science and Engineering Through Cyberinfrastructure* [18] was created by the Blue-Ribbon Advisory Panel on Cyberinfrastructure of the NSF in 2003.
- **DigiCult** – *The Future Digital Heritage Space: An Expedition Report* [6] was published as a DigiCULT thematic issue in 2004.
- **Erpanet** – *Electronic Resource Preservation and Access Network*[2] was a European Commission funded project which ran from 2001 until 2004.
- **Warwick** – *Digital Curation and Preservation: Defining the research agenda for the next decade* [1] reports on the Warwick workshop held in 2005.
- **DRR** – *Digital Repositories Roadmap - Looking Forward* [29] by Rachel Heery and Andy Powell, 2006.
- **OSI**– *E-Infrastructure Strategy for Research: Final Report from the OSI Preservation and Curation Working Group* [44] by Neil Beagrie, 2007. Note: Part of this report is based on the findings of [1], which are not repeated here.

As other Web resources, blogs are not immune from decay or loss. Many blogs that described major historic events, which took place in the recent past, have already been lost (Chen, 2012). Another example that justifies preservation initiatives is the account of disappearing personal diaries. Their loss is believed to have implications for our cultural memory (O'Sullivan, 2005). The dynamic nature of blogging platforms suggests that existing solutions for preservation and archiving are not suitable for capturing blogs effectively. However, blog preservation is not a trivial task.

Among the few studies that raise the need for blog archiving and the potential impact of blog loss is the work by O'Sullivan who highlights the archival potential of blog based diaries and the consequences of losing those. Yet, the review paper by Chen demonstrates that little attention is given to the issue of blog preservation and archiving. Existing solutions available for archiving Web content are limited when applied to archiving the Blogosphere.

PANDORA[47] is the Web Archive of the National Library of Australia. It is considered to have been the first to make a step towards blog preservation in 2004. However, the preservation case was limited to a single blog. The library increased the number of preserved blogs to twelve by April 2011. A more recent approach from ArchivePress[48] allowed coverage of a larger domain. The solution, developed by Pennock and Davis (2009), provided a mechanism for institutions to collectively harvest blog content. They used WordPress Open Source software and RSS feeds to archive parts of blogs believed to be of primary importance and re-use value. The differences in specifications of feed formats and diversity of their implementations impose specific restrictions or challenges. For instance, the content of entries captured from web feeds may be truncated. Consolidating or choosing from multiple web feeds exhibited on a single blog, for instance, may require additional effort.

[47] http://pandora.nla.gov.au/
[48] http://archivepress.ulcc.ac.uk/

The main purpose of the overview presented here was to contextualise the developing interest in web archiving over the years with respect to that in general digital preservation. There have been more comprehensive reviews focusing on the research landscape in digital preservation itself: for example, DigitalPreservationEurope[49] (DPE) produced a review discussing different areas of digital preservation in development at the time of the publication[50]. The table of areas discussed and keys for the research initiatives discussed are included in Table 2.1-2 and Table 2.1-3, respectively, for easy reference. Another noteworthy review is that produced by the Technische Universität Wien (TUW) which discusses EU funded projects[51].

2.2 Relevant Projects and Initiatives

The digital preservation research landscape has been changing actively in the last two decades. These have consisted of a range of collaborative networks (Figure 2.2-1), projects with a focus on creating planning and characterisation tools, infrastructure and conducting experiments (Figure 2.2-2), projects that relate to web preservation (Figure 2.2-3) and initiatives with specific areas of focus (Figure 2.2-4).

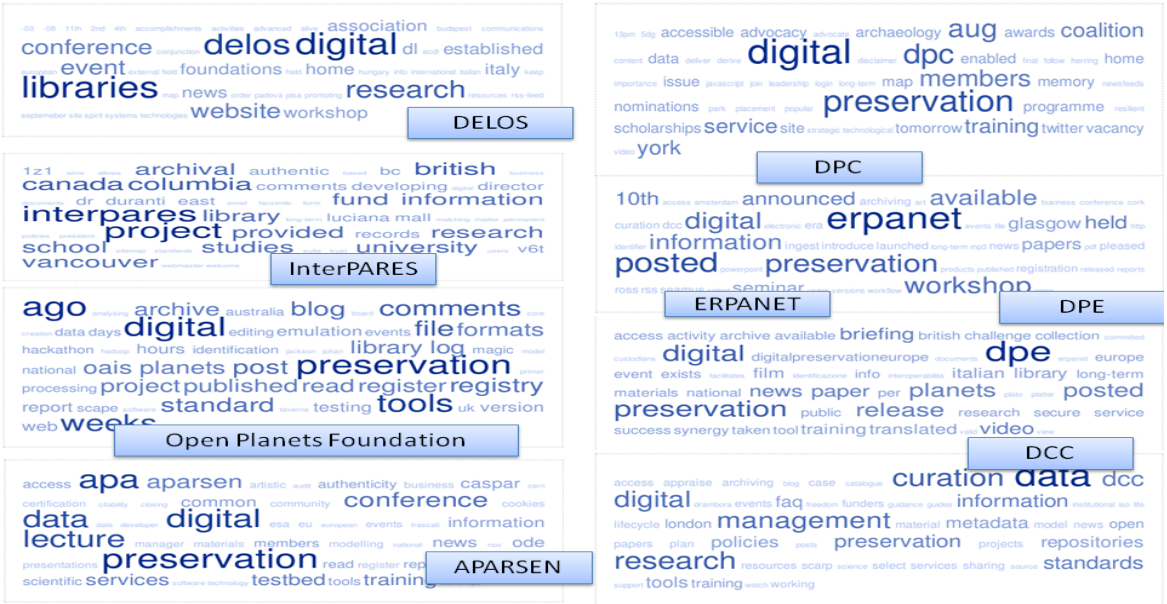


Figure 2.2-1 Collaborative knowledge networks in digital preservation and their word clouds (clouds generated by TagCrowd at http://tagcrowd.com/).

For example, the technology watch papers from the Digital Preservation Coalition[52] (DPC) and the briefing papers and manual chapters at the Digital Curation Centre[53] (DCC) have been invaluable in providing insight into the best practices that have been developed over the years with respect to the adoption of best quality formats, schemas, standards and management practices. Initiatives such as Electronic Resource Preservation and Access Network[54] (ERPANET) and DigitialPreservationEurope[55] (DPE) and served to provide central locations for aggregating and managing knowledge and resources related to digital preservation. Likewise, networks such as

[49] http://www.digitalpreservationeurope.eu/
[50] http://www.digitalpreservationeurope.eu/publications/reports/dpe_research_roadmap_D72.pdf
[51] http://www.ifs.tuwien.ac.at/~strodl/paper/Report%20-%20Research%20on%20Digital%20Preservation.pdf
[52] http://www.dpconline.org/
[53] http://www.dcc.ac.uk
[54] http://www.erpanet.org/
[55] http://www.digitalpreservationeurope.eu/

DELOS[56], International Research on Permanent Authentic Records in Electronic Systems[57] (InterPARES), and more recently, Open Planets Foundation[58] (OPF) and Alliance for permanent Access to the Records of Science in Europe Network[59] (APARSEN), have contributed to the creation of synergies between digital preservation communities.

Many projects, in parallel have worked towards realising the concrete infrastructure necessary for implementing preservation actions and measuring preservation performance. For example, Lots of Copies Keep Stuff Safe[60] (LOCKSS) has fostered libraries to preserve their content by comparing the material across several copies made available to a central system, and Making of America II[61] have been creating best practice standards for description and transmission of metadata, an example of which is the Metadata Encoding and Transmission Standard[62] (METS). The project, Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval[63] (CASPAR), has produced much work in developing the tools and practices to facilitate the implementation of the reference model for an Open Archival Information System[64]. Projects such as Preservation and Long-term Access through Networked Services[65] (PLANETS) and Scalable Preservation Environments[66] (SCAPE) have contributed the environmental infrastructure for experimenting with digital objects (e.g. in the context of format identification, characterisation, migration and emulation) for preservation planning.

Notable contributions of projects like SCAPE, especially for web archiving projects is that they have done much to meet the challenges of scalability with respect to preservation actions and automated processes by integrating methods such as HADOOP/MapReduce[67] to handle *big data* to increase system performance and speed. However, their work to date have been based on material already stored in an archive (e.g. Australia's PANDORA archive), and, as far as complexity is concerned, they only examine heterogeneity of the digital object types: not so much the complexity introduced in light of the various processes that will have to be threaded together. They also do not seem to be considering other distributed computing methods other than HADOOP batch processing (for example, stream processing approaches such as Storm[68] and other approaches that allow more complex processes[69]).

[56] http://www.delos.info/
[57] http://www.interpares.org/
[58] http://www.openplanetsfoundation.org/
[59] http://www.alliancepermanentaccess.org/
[60] http://www.lockss.org/
[61] http://sunsite.berkeley.edu/moa2/
[62] http://www.loc.gov/standards/mets/
[63] http://www.casparpreserves.eu/
[64] http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683
[65] http://www.planets-project.eu/
[66] http://www.scape-project.eu/
[67] For a great synopsis of HADOOP, see http://radar.oreilly.com/2011/01/what-is-hadoop.html
[68] https://github.com/nathanmarz/storm/wiki/Tutorial
[69] See the discussion here: http://highscalability.com/blog/2010/3/30/running-large-graph-algorithms-evaluation-of-current-state-o.html

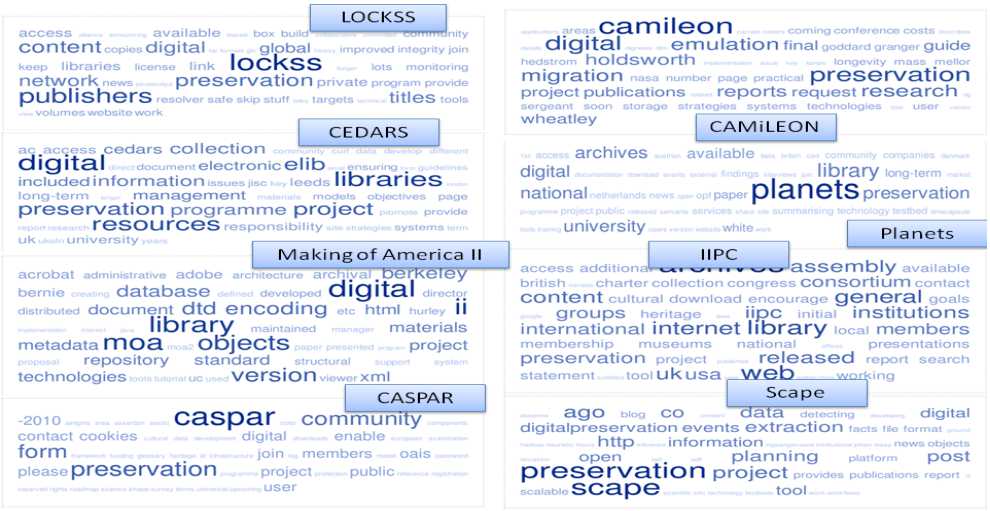


Figure 2.2-2 Initiatives with focus on creating tools, infrastructure and experiments and their word clouds (clouds generated by TagCrowd at http://tagcrowd.com/).

Recent projects in web archiving and web information management (such as ARCOMEM and LAWA) have adopted methods in handling *big data* as a necessary approach to dealing with the ever increasing large volume of web information. These latter initiatives, however, put less emphasis on enabling evaluating preservation processes to be used within the archive, placing more effort on developing methods selection and appraisal of material to be included in the archive.

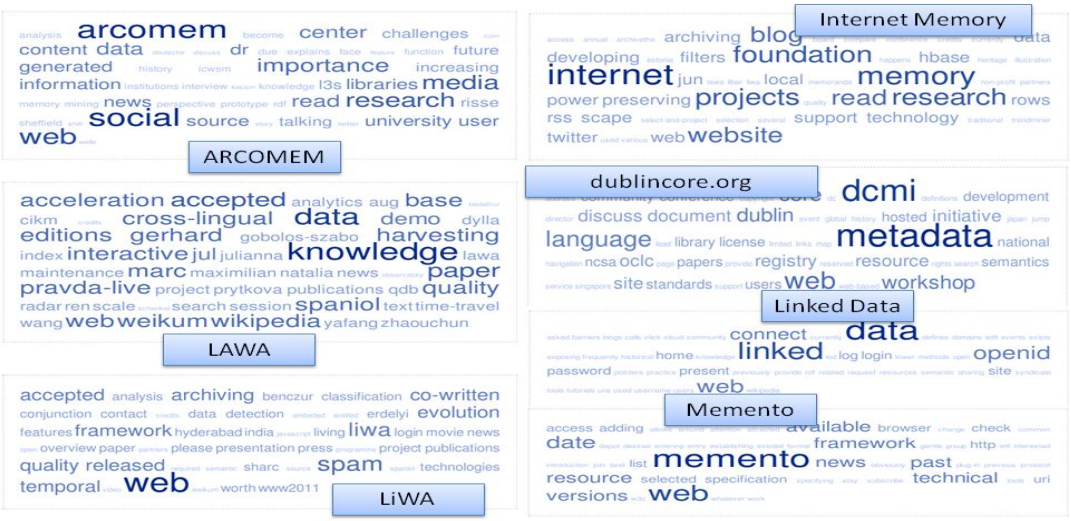


Figure 2.2-3 Initiatives related to web preservation and their word clouds (clouds generated by TagCrowd at http://tagcrowd.com/).

Other initiatives, such as Dublin core metadata[70] (first developed in 1995) created metadata standards for information over the web, while Linked Data[71], and Memento[72] approach web preservation by way of preserving the links to resources and the inter relationships between pieces of data through exposing them publicly as RDF triples and instances on time gateways, respectively.

[70] http://dublincore.org/
[71] http://linkeddata.org/
[72] http://mementoweb.org/

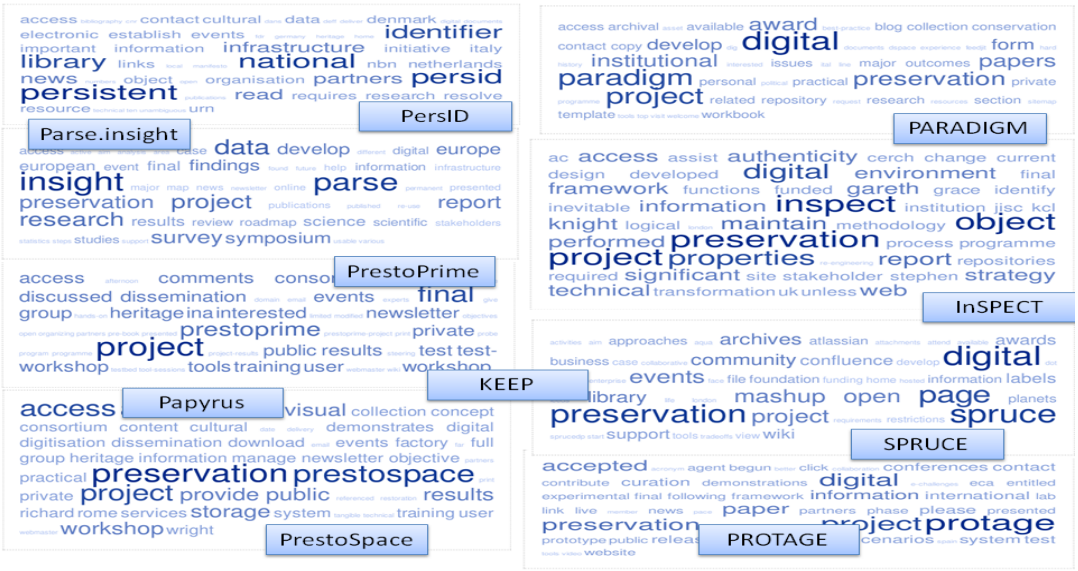


Figure 2.2-4 Initiatives with areas of focus and their word clouds (clouds generated by TagCrowd at http://tagcrowd.com/).

There have been other projects in focused areas, such as those that support, the preservation of audio visual objects (for example, PrestoPrime[73] and PrestoSpace[74]), the creation of persistent identifiers (e.g. PersID[75]), investigations into significant properties (such as InSPECT[76]), the development of intelligent software agents (cf. PROTAGE[77]), the organisation of mashup events for a community problem-driven solution to preservation (for instance, projects such as SPRUCE[78]) and the advancement of emulation approaches to preservation (projects such as KEEP[79]). In particular, the InSPECT project was one of the more recent projects that introduced a strong component addressing the needs of the community as an approach to defining the significant properties of a digital object. The XCL project[80] also developed approaches to significant properties with emphasis on format characterisation.

The problems considered in this report is most comparable to SCAPE with respect to its aim in trying to enable the implementation and testing of preservation processes with respect to materials consisting of very large collections of web pages. However, there are some notable differences. These have been presented in Figure 2.2-5.

With respect to characterisation of weblog features to determine significant properties, on the micro level of digital object types embedded in the weblog, we adopt approaches similar to that proposed in the InSPECT project (Section 3.4). We deemed InSPECT to be most appropriate due to their focus on community needs. Other approaches such as that used in the eXtensible Characterisation Language (XCL) project were also considered. However the adoption of results from the XCL project, had notable drawbacks: a) the approach is primarily designed to combat format obsolescence, that is, its framework concentrates on format characterisation and has not been tested to express inter relationships between objects, a prominent aspect of weblogs as identified by the WP2 work on the BlogForever data model (Stepanyan et al. 2011), b) at the time of this report, it was felt that the approach had not been extensively tested on the basis of meeting end-user requirements, and c) the characterisation did not seem to extend to profiling the designated

[73] http://www.prestoprime.org/
[74] http://www.prestospace.org/
[75] http://www.persid.org/
[76] http://www.significantproperties.org.uk/
[77] http://www.protage.eu/
[78] http://wiki.opf-labs.org/display/SPR/Home
[79] http://www.keep-project.eu/ezpub2/index.php
[80] http://planetarium.hki.uni-koeln.de/planets_cms/about-xcl

community which was deemed a necessity for social media content created by and for blogging communities.

Our approach, however, in contrast to InSPECT, extends to consider significant properties on the macro level of inter-connected weblog components (Section 3.3). And, further, we consider the way different communities use the micro level objects within the macro level context.

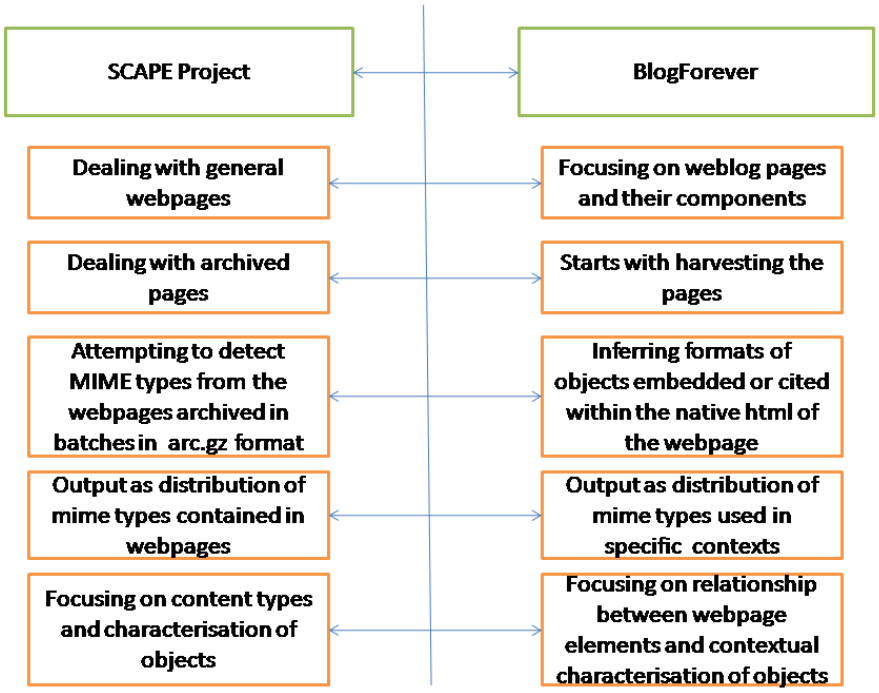


Figure 2.2-5 Comparison of preservation objectives: SCAPE versus BlogForever.

In terms of cognizance of community needs, we might be compared to the SPRUCE project, but, unlike the SPRUCE project that relies on the agile development of solutions for each arising preservation problem within selected communities, we aim to develop a community profiling approach that can pre-empt problems that might arise within communities based on technical conventions, network structures and information sharing behaviour.

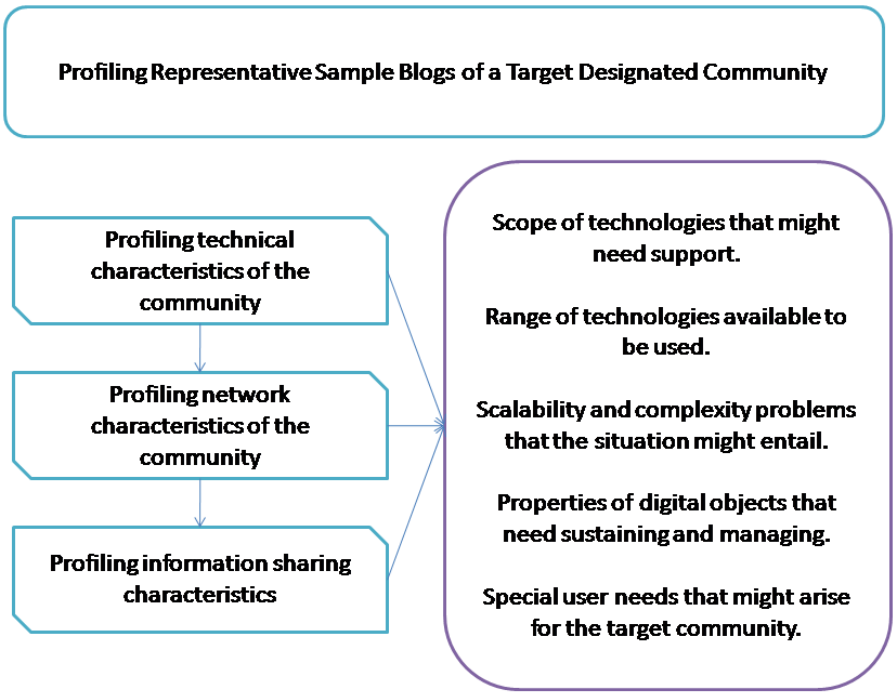
While the emulation approach in active development within projects such as KEEP may potentially provide a solution for the complex nature of weblogs, the application of emulation environments to webpages and websites characterised by heterogeneous digital objects and inter-connections, as well as, varying requirements on web browsers and operating systems (see Chapter 4) is premature. Migration as a strategy could be viable but the complexity of inter-connections, the heterogeneity of digital object types and formats, and sheer volume of the weblogs to be harvested and updated makes the scalability of such an approach also questionable.

Our contribution here, that distinguishes it from previous projects, is threefold:

1. We combine automated feature extraction with object analysis and user requirement analysis to develop a robust preservation strategy based on the characterisation of weblogs from a target weblogging community. The feature extraction, in particular, leads to the development of criteria for building datasets for user validation and alternative preservation strategies testing, which is representative of the complexities of weblogs coming from a target designated community (Table 2.2-1).

2. We propose a service-requirements based practical implementation of the repository that reflects OAIS-like functions, thereby transforming the high-level concepts of the OAIS to a transferable workflow for weblog repositories.
3. We propose features that could be added to the repository design that would result in a *social repository* for social network data repository. This could help to alleviate the scalability problem that will only get worse in the future as both volume and complexity of web information increases.

Table 2.2-1 Defining features of weblogs.



In this chapter, we presented a brief overview of the research in digital preservation likely to be relevant to weblog preservation. In subsequent chapters we will return to some of these projects for a more discussion. More specifically, for example, in Section 3.3, we will return to a discussion of the project InSPECT to discuss significant properties of weblogs. We will also return to a discussion of encoding standards such as METS in our discussion of metadata schemas we will be using to support the preservation of weblogs in Chapter 5. The SCAPE project will be discussed along with some work produced from the Internet Preservation Consortium (IIPC), as well as, recent reports on format identification within the UK Web Archive[81], in Chapter 4, where we discuss the scalability and complexity of implementing preservation processes. We will also return to further discussion of the results produced by community-driven projects such as SPRUCE, in Chapter 7, where we will expand on the necessity to get the community involved in the web archiving project.

[81] http://britishlibrary.typepad.co.uk/webarchive/2012/08/analysing-file-formats-in-web-archives.html

3 Weblogs

In discussing weblogs, it is easy to make the assumption that there exists an unambiguous established notion of weblogs common to all. While there seems to be an intuitive feeling for weblogs as online publishing channels defined by the reverse chronological order of posts contributed by authors, any other attempt towards a formal definition can easily lead to counter examples.

For example, weblogs, unlike Facebook or Twitter, do not have a centralised uniform platform leading to similar technical output. Instead, they emerge through the use of several different platforms (e.g. WordPress[82]) the identity of which are often hidden and the manifestation of which are even highly customised. This makes the task of preserving weblogs that much harder. Nevertheless, in this Chapter, we try to bring some immediate sense to what a blog is by drawing on the BlogForever: D2.1 Weblog Survey, BlogForever: D2.2 Weblog Data Model (Stepanyan et al. 2011), BlogForever: D4.1 User Requirements and Platform Specification Report (Kalb et al. 2011) and previous studies.

First we present an overview of how bloggers and blog readers might perceive blogs and what types of digital objects and formats might be found within weblogs (Section 3.1). Then we summarise the data model constructed as part of WP2 deliverable D2.1 to show how these objects are situated within the larger structure of a weblog (Section 3.2). From this we derive the notion of four archive records that will be supported by the BlogForever repository (Section 3.2.6). In Section 3.3, we revisit the user requirements examined within WP4 deliverable D4.1 to try to narrow down the significant components of the data model. This will define which components will be the main targets of preservation planning. Finally, we discuss results from previous studies to address the significant properties of the object types identified in Section 3.1 (Section 3.4) and conclude with a remark on the next steps (Section 3.5).

3.1 Weblog Survey

The project report BlogForever: D2.1 Weblog Survey feeds into all aspects of the current task: the technical survey directly impacts the data model and related object analysis, and the perceived value of blogs expressed by survey participants provides insight into potentially significant properties of blogs.

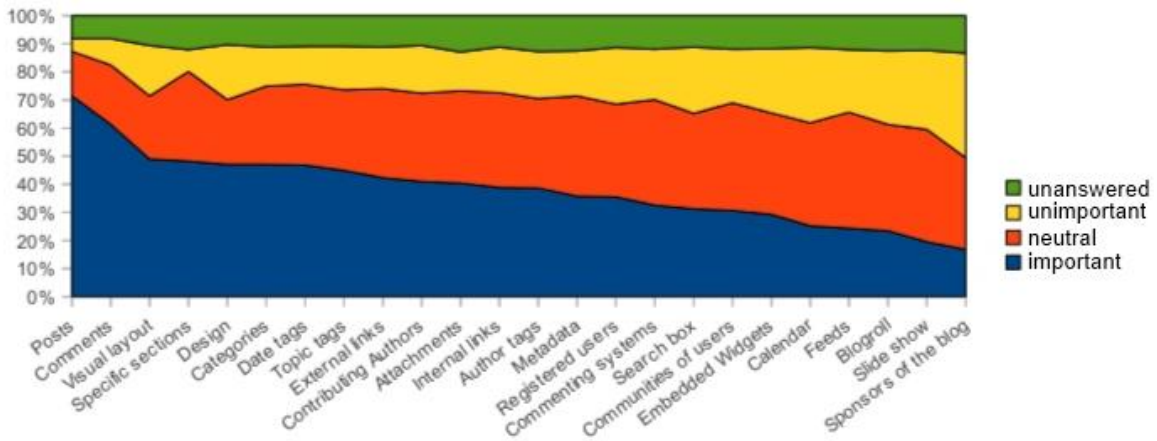


Figure 3.1-1 Value of blog components perceived by survey participants.

[82] http://www.wordpress.com

The graph in Figure 3.1-1 illustrates the values assigned to different elements within a blog, by those who took the weblog survey. The elements are ordered according to perceived "importance". In the original survey of D2.1, this was displayed in Table 30, using five categories of importance from "very unimportant" to "very important". Here we also display the proportion of participants for whom there was no answer (labelled "unanswered" and the area coloured green) and we merge "very unimportant" with "unimportant" and "very important" with "important", resulting, altogether, in four categories (beginning from the top of the graph, these categories are "unanswered", "unimportant", "neutral", and "important"). The figure shows that visual characteristics of blogs (such as "visual layout" and "design"), are also perceived as valuable to many participants of the survey. The stakeholder groups in terms of age (Table 3.1-1) and background (
Table 3.1-2) also vary across a wide spectrum, and, likewise, the target audience (Table 3.1-3) and motivations (Table 3.1-4) seem to vary widely as well.

Table 3.1-1. Age of survey participants (Table 9, deliverable D2.1).

Table 9 – Authors by age group

AgeGroup	Responses	%
25 - 34	131	25.6
35 - 44	106	20.7
18 - 24	82	16
45 - 49	49	9.6
Under 18	47	9.2
55 - 64	42	8.2
50 - 54	37	7.2
Over 65	18	3.5

Table 3.1-2. Background of survey participants (Table 6, deliverable D2.1).

Table 6 – Authors by education & employment

EducationEmployment	Responses	%
In paid employment	207	40.4
In full-time education	122	23.8
Freelancer	55	10.7
Self-employed	49	9.6
Home carer	27	5.3
Other	52	10.2
Total	*512*	*100*

Table 3.1-3. Target of blog as specified by survey participants (Table 14, deliverable D2.1).

Table 14 – Main audience of blogs

MainAudience	Responses	%
General Public	*306*	59.8
Family and Friends	*207*	40.4
Myself	*165*	32.2
Colleagues and Professional Peers	*164*	32.0
Students	*91*	17.8
OTHER BLOGGERS	*6*	0.11
Other	*45*	8.8

Table 3.1-4. Motivation for blogging (table 24, deliverable D2.1).

Table 24 – Motivations for maintaining blogs

Values	Responses	%
Personal	408	79.7
Information sharing	311	60.7
Discussion of topics	250	48.8
Mostly for myself	218	42.6
Create an online presence	195	38.1
Professional	181	35.4
Entertainment	165	32.2
Record of activities or events	158	30.9
Mostly for my audience	139	27.1
Organise / promote / support an activity	119	23.2
Promote teaching and learning	108	21.1
Commercial	47	9.2
Manage a project	39	7.6
To target markets or communities	39	7.6
Manage a conference	12	2.3
Other	33	6.4

Just as the perceived value of blog components, as well as the background and the motivation of stakeholders vary, the types of digital objects vary also. The common digital object formats found within weblogs were identified as part of BlogForever deliverable D2.1: *Survey Implementation Report*, Section 5: Technology Used by Current Blogs. These are categorised here into six types:

A. Structured Text
B. Image
C. Audio
D. Moving Images
E. Documents
F. Executables

These digital object types are explained in the following subsection with example formats found as part of the weblog survey presented using tables. The tables list formats, some of their common file extensions, object type, mime type and the reference in D2.1 where the format is identified.

While there is other categorised content (see Section 3.2 and Appendix F, page 228) found in weblogs such as links and tags, the above six types are highlighted as they are expected to involve non-trivial support with respect to metadata.

3.1.1 Digital Object Type: Structured Text

Definition: a plain text-based object (not the same as a Document – see Section 3.1.3). The key characteristic that distinguishes structured and unstructured text is the presence of mark-up that provides additional information about the interpretation of text.

> *"The key characteristic that distinguishes structured and unstructured text is the presence of mark-up that provides additional information about the interpretation of text. The central premise of the Performance model is the distinction between the raw, uninterpreted data, defined as the Source, and the interpretation of the data as a Performance. Although this is a useful metaphor, its application for structured text documents will vary, as distinguished by the content type and the rendering method. During the analysis it was recognized that, when applied to certain types of structured text (e.g. XML documents that do not possess associated instructions on the preferred method of recreation), the Performance Model metaphor is unhelpful unless a distinction between the Source and Performance can be made. Many types of structured text may be 'performed' using several different methods. To illustrate, an XML-encoded text may be presented to the user as an RSS feed, processed and converted to an audio stream, or represented in several XHTML-compliant web pages that contain different types of information."* [83]

Table 3.1-5 Example File Formats for the Structured Text Object Type

File format identified	File extension
Hyper Text Markup Language	HTML, HTM
eXtensible HyperText Markup Language	XHTML, XHT
Extensible Markup Language	XML
PHP Script Page	PHP
HTML File Containing Server Side Directives	SHTML
Cascading Style Sheet	CSS

3.1.2 Digital Object Type: Image

Definition: Digital (still) images are non-moving representations of visual information. That is, still images that convey their meaning in visual terms, e.g. photographs, posters, diagrams, drawings. The AHDS study considers both the familiar raster image and the perhaps less well known vector image. The former include the products of digital photography and scanning with file formats such as TIFF and JPEG. The latter is considered less when thinking of digital images, but a large volume of digital content is created including maps, drawings, and the almost ubiquitous PDF file. (From AHDS Digital Images Archiving Study, 2006).

[83] From The InSpect final report (2009), http://www.significantproperties.org.uk/inspect-finalreport.pdf

Table 3.1-6 File Formats for the Image Object Type

File format identified	File extension
Portable Network Graphics	PNG
Graphics Interchange Format	GIF
Bitmap	BMP
JPEG	JPG
Scalable Vector Graphics	SVG

3.1.3 Digital Object Type: Document

The DELOS report [84] used the term "Document-like" as part of their typology of file formats. As part of their definition, they identified:

> *"**Documents created permanently**: the content is permanently stored inside these documents. Both the structure and content are usually defined at the moment the document is created, by using tools that work on the abstract internal representation of the document. Typical examples of these formats are the PDF format or the Microsoft Word format."*

Table 3.1-7 File Formats for the Document Object Type

File format identified	File extension
MS Word for Windows Document	DOC
MS Office Open XML	DOCX
OpenDocument Text	ODT
Portable Document Format	PDF
Plain Text File	TXT
MS Excel Workbook	XLS
MS Excel for Windows	XLSX
OpenDocument Spreadsheet	ODS
MS PowerPoint	PPT
MS PowerPoint for Windows	PPTX
OpenDocument Presentation	ODP

3.1.4 Digital Object Type: Audio

Definition: Sound resources include digitally recorded audio and digitised versions of analogue sound files[85].

Table 3.1-8 File formats for the audio object type

File format identified	File extension

[84] See DELIVERABLE REFERENCE NUMBER: WP6, D6.3.1, File formats typology and registries for digital preservation (2004), http://www.dpc.delos.info/private/output/DELOS_WP6_d631_finalv2%285%29_urbino.pdf

[85] From Digital Moving Images and Sound Archiving Study, AHDS (2006), http://www.ahds.ac.uk/about/projects/archiving-studies/moving-images-sound-archiving-final.pdf

MPEG 1/2 Audio Layer 3	MP3
Waveform Audio	WAV

3.1.5 Digital Object Type: Moving Image

Definition: Moving image resources include streaming video (e.g. digital television broadcasts), the outputs of moving image capture devices, such as consumer and professional video cameras, and digitised versions of analogue video formats[86].

Table 3.1-9 File formats for the moving image object type

File format identified	File extension
MPEG-1 Video Format MPEG-2 Video Format	MPEG, MPG
Audio/Video Interleaved Format	AVI
QuickTime	MOV
3GPP Audio/Video File	3GPP
Macromedia FLV	FLV

3.1.6 Digital Object Type: Executable

Definition: These are the executable components of a complex object, such as a CD-ROM or Web document. These executables perform certain operations within the digital object. They are not the software stated in system requirements, though they may be supported by it.

Table 3.1-10 File formats for the executable object type

File format	Common file extensions	Digital object type
Postscript	AI EPS EPSF PS	Executable
Base64-encoded bytes	MM MME	Executable
UNIX tar file, Gzipped	GZ TGZ Z ZIP	Executable
Compressed archive file	ZIP	Executable
Gzip compressed archive file	GZ	Executable
Tape Archive Format	TAR	Executable
Zip Format	ZIP	Executable
Executable file	EXE DLL MSI	Executable
XPInstall	XPI	Executable
Atom Syndication Format feed	ATOM	Executable
Really Simple Syndication feed	RSS	Executable

[86] From Digital Moving Images and Sound Archiving Study, AHDS (2006), http://www.ahds.ac.uk/about/projects/archiving-studies/moving-images-sound-archiving-final.pdf

File format	Common file extensions	Digital object type
Resource Description Framework	RDF	Executable
Really Simple Discovery	RSD	Executable
JavaScript	JS	Executable

3.1.7 File Formats From the Weblog Survey

The lists in Table 3.1-11, Table 3.1-12, Table 3.1-13 summarises the format types in terms of frequency as discovered as part of the weblog survey. This will be expanded upon within Chapter 4, when we discuss, in more detail, the object types and formats used within different blogging communities.

Table 3.1-11 Formats occurring frequently within weblogs

File format identified	Common file extensions	Digital object type	mime type	D2.1 Report reference
Hyper Text Markup Language	HTML, HTM	Structured text	text/html	5.2.2
eXtensible HyperText Markup Language	XHTML, XHT	Structured text	application/xhtml+xml	5.2.2
Extensible Markup Language	XML	Structured text	text/xml	5.2.2
PHP Script Page	PHP	Structured text	text/html	5.2.2
HTML File Containing Server Side Directives	SHTML	Structured text	text/html	5.2.2
Cascading Style Sheet	CSS	Structured text	text/css	5.2.3
Portable Network Graphics	PNG	Image	image/png	5.2.3
Graphics Interchange Format	GIF	Image	image/gif	5.2.3
Bitmap	BMP	Image	image/bmp	5.2.3
JPEG	JPG, JPEG	Image	image/jpeg	5.2.3
Scalable Vector Graphics	SVG	Image	image/svg+xml	5.2.3
MPEG 1/2 Audio Layer 3	MP3	Audio	audio/mpeg	5.2.8
Waveform Audio	WAV	Audio	audio/x-wav	5.2.8
MPEG-1 Video Format MPEG-2 Video Format	MPEG, MPG	Moving images	video/mpeg	5.2.8
Audio/Video Interleaved Format	AVI	Moving images	video/x-msvideo	5.2.8
Quicktime	MOV	Moving images	video/quicktime	5.2.8
3GPP Audio/Video File	3GPP	Moving images	video/3gpp	5.2.8
MS Word for Windows Document	DOC	Document	application/msword	5.2.8
MS Office Open XML	DOCX	Document	application/vnd.openxmlformats-officedocument.wordprocessingml.document	5.2.8
OpenDocument Text	ODT	Document	application/vnd.oasis.opendocument.text	5.2.8
Plain Text File	TXT	Document	text/plain	5.2.8

File format identified	Common file extensions	Digital object type	mime type	D2.1 Report reference
MS Excel Workbook	XLS	Document	application/vnd.ms-excel	5.2.8
MS Excel for Windows	XLSX	Document	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	5.2.8
OpenDocument Spreadsheet	ODS	Document	application/vnd.oasis.opendocument.spreadsheet	5.2.8
MS PowerPoint	PPT	Document	application/vnd.ms-powerpoint	5.2.8
MS PowerPoint for Windows	PPTX	Document	application/vnd.openxmlformats-officedocument.presentationml.presentation	5.2.8
OpenDocument Presentation	ODP	Document	application/vnd.oasis.opendocument.presentation	5.2.8

Table 3.1-12 Formats that occur infrequently within weblogs

File format identified	File extension	Type	mime type	D2.1 Report reference
Wireless Bitmap	WBMP	Image	image/vnd.wap.wbmp	5.2.3
WebP	WEBP	Image		5.2.3
Tagged Image File Format	TIFF, TIF	Image	image/tiff	5.2.3
Macromedia FLV	FLV	Moving images	video/x-flv	5.2.3
Macromedia Flash	SWF	Flash	application/x-shockwave-flash	5.2.3
MS Access Database	MDB	Database	application/msaccess	5.2.8
	CCBD (?)	Database		5.2.8
OpenDocument Database Format	ODB	Database		5.2.8

Table 3.1-13 Formats that were not found in the weblog survey

File Format Identified	File Extension	Type	Mime Type	D2.1 Reference
MS Access Database	MDB	Database	application/msaccess	5.2.8
	CCBD (?)	Database		5.2.8
OpenDocument Database Format	ODB	Database		5.2.8

3.1.8 Next Steps

The Weblog Survey described in BlogForever deliverable D2.1 was carried out to examine whether selected common formats were found within the blogosphere. There was no attempt to conduct an exhaustive search of what formats are being used within selected communities. The survey was carried out to detect isolated instances of technology use: that is, there was no examination of the environment (technical and social) in which these technology uses arise.

In the next sections, the BlogForever data model (from deliverable D2.2) will be summarised, partially illuminating the contexts within which these different object types arise. This will be combined with the BlogForever user requirements and platform specification (deliverable D4.1) to narrow down the characterising features, the *significant properties of a blog.*

However, given the broad spectrum of blog user age, educational background, and perception regarding blog element value (Section 3.1.1), from a preservation perspective, to define significant properties of weblogs that would serve a designated community, we contend that it is necessary to develop a more explicit notion of community. As a response we carry out a large scale study of web pages within the blogosphere to be compared to other webpages along technological, topical and social network dimensions (see Section 3.4 and Chapter 6).

3.2 Weblog Data Model and its Properties

This section draws from the investigation conducted as part of the Work Package Two (WP2). More specifically, it focuses on the inquiry into the semantics of blogs as part of the Task 2.2. It outlines the results and approach used for developing a blog data model (Stepanyan, K. et al, 2011) and discusses it in the preservation context.

3.2.1 Introduction

WP2 consisted of three primary tasks: *Task 2.1, Conduct Weblogs Survey; Task 2.2: Explore Weblog Semantics; and Task 2.3: Investigate Weblog Data Extraction.* The focus of this chapter bounded to the Task 2.2.

One of the outcomes submitted as part of the T2.2 task was the deliverable D2.2 (ibid) that proposed a blog data model informed by a set of inquiries. The data modelling took into consideration user views from the earlier conducted online survey and recommendations from the theoretical inquiry into network analysis, supplemented by the inquiries into such as, the existing conceptual models of blogs, the data models of Open Source blogging systems and data types identified from an empirical study of web feeds.

3.2.2 Data Modelling

Data Modelling is considered to be an integral phase for designing and developing data systems. Although essential to a design process, the methods for developing data models vary widely. The differences across data modelling practices are reflected in the principles/paradigms of modelling, approaches and methods used, as well as representational notations and standards.

Most frequently, data modelling is conducted by defining the requirements. The rationale behind drawing a set of requirements is to ensure that the data model addresses these requirements for the solutions that are being developed. The requirement definition stage, as suggested by Ponniah (2007), may include interviews, group sessions, documentation, change management and so on.

However, the primary requirements of the project have already been defined and agreed as part of the project agreement[87].

Given the generic requirements of the task of data modelling was to explore the structure of blogs to be able to accommodate a range of blogs and their properties. Hence, the proposed blog data model was developed in a number of consecutive phases. Each of the phases contributed to the process of informing the development of the proposed model.

For the purposes of the BlogForever project, conceptual and more detailed logical information levels have been chosen for representing the proposed data model (Stepanyan, 2011). The decision was based on the necessity to provide both a high level view as well as the more detailed one.

3.2.3 Methods Used

This structured approach required each of the development cycles to inform the process of data modelling leading to the review and refinement of the data model. The cycles included the following consecutive steps.

1. An insight into the database structure of open source blog systems.
2. A retrospective view on an earlier conducted online user survey to identify important aspects and types of blog data to be preserved
3. A retrospective view on the technologies and standards used within the Blogosphere.
4. Suggestions derived from an earlier inquiry into the recent developments and prospects for analysing networks and dynamics of blogs.
5. An inquiry into blog structure based on evaluation of 2,695 blog feeds.
6. An insight into blog APIs.
7. Consultation exercise from a blog service provider Phaistos.

Therefore, the development of the data model was based on understanding the concepts that were identified as integral to blogs and the relationships among these.

3.2.4 Outline of the Data Model

It is evident that blogs are multi-faceted entities that may require a range of different data structures to be put in place. However, it is also apparent that most of the blogs share common features and a general outline. It is therefore possible to develop a generic and simple data model that could suffice the preservation of the basic components of the blogs. This basic model – referred here as the core model – can then be extended to ensure their integrity and the requirements of successful preservation.

The components of the core model were identified by looking into user views on blogs, their database structure, the structure of their web feeds and types of data distributed by them. By looking into both technical specification as well as a summary of user perceptions it was possible to identify most prominent conceptual components of blogs referred to in the model as entities.

These components were further studied in order to identify and describe their properties. The properties of these components constitute the data they carry and metadata used to describe them. These properties have been collected and analysed before integrating them into the data model. Once the data type and association with the entities were identified, the properties have been integrated into the data model.

The following section summarises the data model and its properties. The detailed report about the inquiries used for developing the data model is available in the D2.2 report (Stepanyan, 2011).

[87] *Grant Agreement Annex I - Description of Work (DoW).*

3.2.5 Blog Core

Early in the process of inquiry, it becomes clear that an established vocabulary is associated with blogs. While the vocabulary at times contains more than one term for referring to the same concept, the use of many terms has been widely accepted. This observation is confirmed at various stages of the conducted inquiries. For example, the review of the existing data models of blogs confirms to the established vocabulary and the use of certain terms. Similar outcomes are revealed after an inquiry into the existing database structures of Open Source blogs. It appears that the concept such as Post, Comment, Page, and Author, appear frequently do describe various sections of websites referred to as blogs. These conceptual entities have been put together to form the core of the blog as described in the data model.

We are using a graphical notation to present the main components of the blogs (Figure 3.2-1). This data model enables storing information carried by and about the above mentioned entities. The interrelation between the identified entities is shown and described by the connected lines. The small triangles indicate the directions of the relationships.

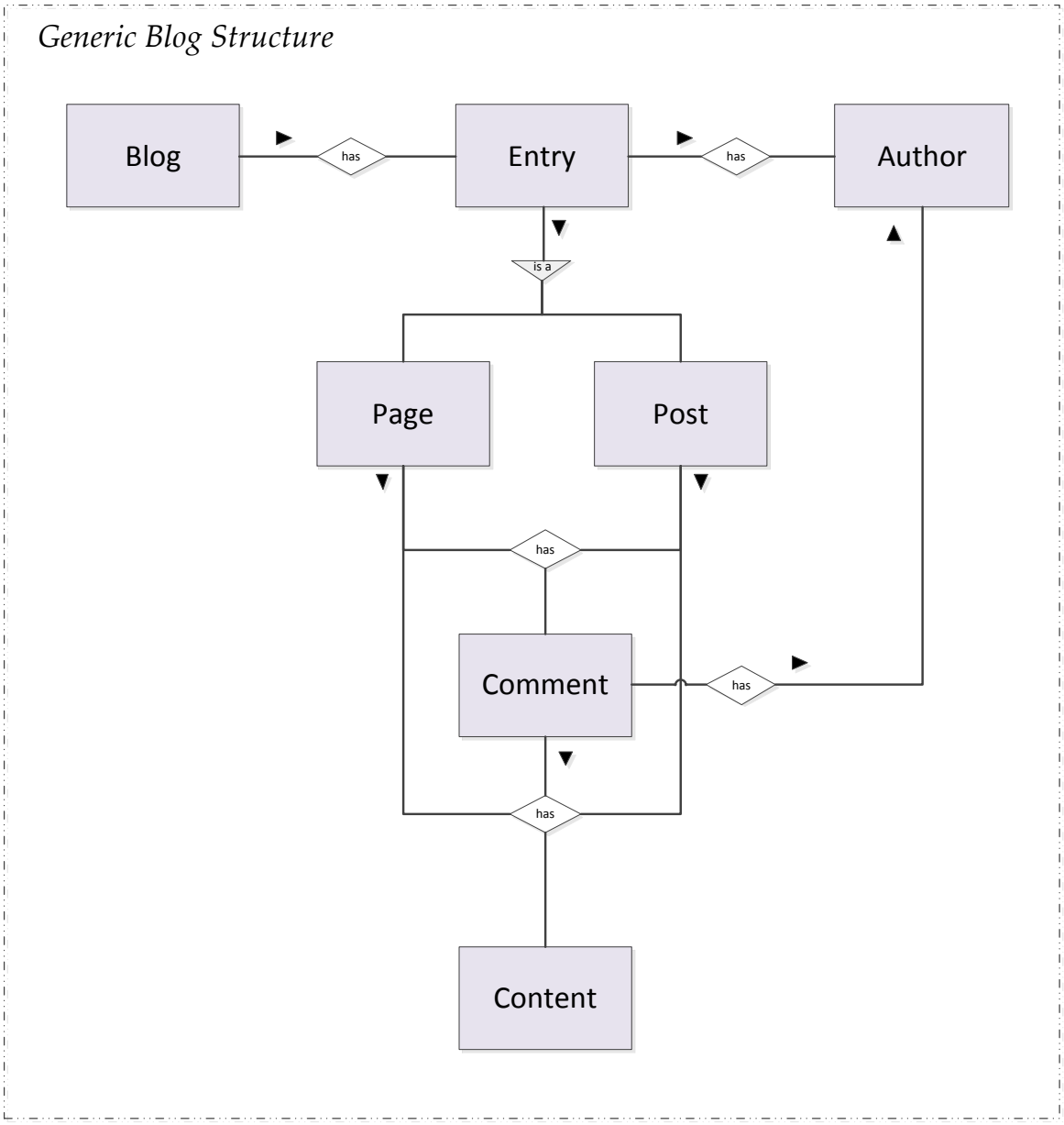


Figure 3.2-1: Blog Core: Data Model

The figure above is presented here to demonstrate the relationships between the major entities of a blog. The model classifies the entities by indicating inheritance (shared properties between entities), the cardinality (demonstrating the types) of the relationships.

This model above demonstrates a high level view of the blog core. However, sets of inquiries mentioned above allowed identifying the properties that can be associated with each of the entities. These properties were collected and integrated into a more detailed view. The vocabulary to describe the properties was further collated. The selected naming was discussed and adjusted when further clarity was needed. Feedback that included the tacit knowledge of partners has also been taken into consideration before developing the data model.

In addition to the inquiries conducted and reported as part of the D2.2 report. The data model was revisited after the completion of the WP4 task that aimed to identify User Requirements for the BlogForever system. The requirements (see Kalb et al. (2011) for details) were identified as a result of interviews conducted with a range of stakeholders. The requirements chosen to be implemented were combined with the existing properties of the data model in an attempt to identify significant properties for preservation. This is further discussed in Section 3.2.7 and Section 3.3 of this report. However, the considerations of the requirements led to updating the model with additional blog properties that would be necessary for providing the services according to the identified requirements. For instance, the time zone in addition to the date properties associated with the Entry or a Comment was added to the data model. Additional information about the date of creation and capture was added.

It is a common practice to anticipate some changes within the data model at the later stages of the project development. After initiating the design of the spider some elements have already been discussed and modified. An internal mechanism for documenting the changes within the data model is already in place.

3.2.6 Records within the Repository

While the data model represents the structure of the data represented in blogs, it is also necessary to identify if blog data can be injected into the repository and, subsequently, presented to the repository user.

Records are information units collected and stored in a repository. Repositories usually contain specific types of records, for instance book, journal or article records. However, in addition to representing physical objects such as printed books, the records can also represent digital material. The collection of records can then be indexed and searched by users. In the context of BlogForever, the records are digital due to the nature of the archived object. Providing meaningful search functionality within a BlogForever repository requires understanding the nature of the objects represented through the stored records.

By looking into the Core Data model we can see that there are a number of prominent entities associated with a blog (i.e. Entry, Comment, and Author). It is likely that users of the BlogForever repository will be willing to search through certain units of information. While using various metadata that describe these units, the outcome of the search should be presented as a set of records that users can access. For instance, a user can be interested in locating a book on the topic of interest, or a simply chapter within a book. Similarly, in the case of blogs, users may wish to locate posts or the entire blog associated with the search term used. Taking into account the possible ways in which BlogForever archive can be used, the following four types of records have been identified: *Blog, Post, Page and Comment.*

Each of the record types can be used for implementing a faceted search functionality, as well as general search by keywords. The keywords entered by the users for searching through the repository can then be compared with the metadata/data stored in association with the records. While keyword search can be based on some complex concepts such as author, the result of the search will be presented as a list of records of the chosen type.

The following sections describe the attributes of the records as presented within the blog data model.

Blog as a Record:
The blog record contains the primary description of the object. The records, contains the name, URL, platform used, etc. Referring to the data model, the record can be described using the following properties.

Table 3.2-1: Description of the Blog Record

Record	Properties	Description
Blog	Title	Title of the blog
	html_title	Contains the title of the HTML head element
	alt_title	Alternate title may include subtitles of the blog or other titles
	alt_title_type	Alternative title type specified the type of the alt_title
	URI	URI of the blog
	status_code	Status code (may reflect whether the blog ceased to exist)
	Language	Retrieved language field, as defined by the blog
	Encoding	Retrieved encoding (character set) field, as defined by the blog
	sitemap_uri	URI of the blog sitemap if exists
	Platform	Platform of the blog powering service, retrieved where available
	platform_version	Versioning information about the platform
	Webmaster	Information about the webmaster where available
	hosting_ip	IP address of the blog
	location_city	Location city based on the hosting details
	location_country	Location country based on the hosting details
	last_activity_date	Date as retrieved from the blog, including time zone
	post_frequency	As retrieved from the blog
	update_frequency	As retrieved from the blog
	Copyright	Notes of copyright as retrieved from the blog
	ownership_rights	Notes of ownership rights as retrieved from the blog
	distribution_rights	Notes of distribution rights as retrieved from the blog
	access_rights	Notes of access rights as retrieved from the blog
	Licence	Licence of the content

Post as a Record:
Posts are entries published by the blog author/s that appear in a chronological order or in categories and are distributed by web feeds. Records Post and Page share most of their properties. While conceptually different, their structure can be seen as very similar. For instance, both Pages as well as Posts can have a name, a unique URL, creation date at so on. Hence, the shared properties are being combined here as Entry (see Table 3.2-2). The properties of the entry are then extended to include the properties relevant for the post only.

Table 3.2-2: Shared Properties for a Post and Page Records

Record	Properties	Description
Entry	Title	Title of the entry
	Subtitle	Subtitle of the entry if available

	URI	Entry URI
	alt_identifier (UR)	A common alternative identifier similar to DOI.
	date_created	Retrieved from the blog or obtained from the date/time crawling, including time zone
	date_modified	Retrieved from the blog or obtained from the date/time crawling, including time zone
	Version	Auto-increment: derived version number (versioning support)
	status_code	Information about the state of the post: active, deleted, updated (versioning support)
	response_code	HTTP response code
	geo_longitude	Geographic positioning information
	geo_latitude	Geographic positioning information
	access_restriction	Information about accessibility of the post
	has_reply	Derived property (also SIOC[88])
	last_reply_date	Derived property (also SIOC), including time zone
	num_of_replies	Derived property (also SIOC)
	child_of	ID of entry parent if available

Table 3.2-3 presents the properties of the Post complementing Entry properties.

Table 3.2-3: Extended Properties of a Post Record

Record	Properties	Description
Post	Type	Custom type of the post if specified (e.g. WordPress): attachment, page/post or other custom type
	posted_via	Information about the service used for posting if specified
	previous_URI	URI to the previous post is available
	next_URI	URI to the next post if available
	Author	See Section Associated Data
	Content	See Section Associated Data

Page as a Record:

While Pages are similar Posts in their properties, their content is not being distributed via web feeds. Furthermore, the pages are not displayed in a chronological order either. However, Pages usually contain relevant information that may describe the Author, and/or provide basic information about the blog etc. Hence capturing the Pages in addition to the Posts is considered important. It has been observed that pages there most of their properties with posts. A different template used for the pages is the only property associated with a Page.

Table 3.2-4: Extended Property of a Page

Record	Properties	Description
Page	Template	Information about the design template if available and if different from the general blog
	Author	See Section Associated Data
	Content	See Section Associated Data

Comment as a Record:

Comments are entries published by others or the author him/herself as a response to the original Page/Post. Unlike Posts or Pages, the Comments appear along with the published entry and provide an opportunity for the readers to voice their views. The control over the publication of the comments is held by the authors/administrators of the blog. The Properties of the Comment are presented in Table 3.2-5.

[88] http://sioc-project.org/ontology

Table 3.2-5: Properties of the Comment

Record	Properties	Description
Comment	Subject	Subject of the comment as retrieved
	URI	URI of the comment if available
	Status	Information about the state of the comment: active, deleted, updated (versioning support)
	date_added	Date comment was added or retrieved, including time zone
	date_modified	Date comment was modified or retrieved as modified, including time zone
	addressed_to_URI	Implicit reference to a resource
	comment_type (UR)	Classification of the comment, i.e. internal or external
	source_URI (UR)	Source of the external service
	source_name (UR)	Name of the external service
	geo_longitude	Geographic positioning information
	geo_latitude	Geographic positioning information
	has_reply	Derived property (also SIOC)
	num_replies	Derived property (also SIOC)
	is_child_of_post	Indicates information about the parent post
	is_child_of_comment	Indicates information about the parent comment
	Content	See Section Associated Data

Other Data Associated with the Record: Content and Author

In addition to the properties discussed along with the records, it is necessary to highlight the existence of other data associated with the records. Most prominent types of data associated with records are: author data and published content. These data, unlike other properties, cannot be described using a single property. For example, authors can have a fist/second name, a username or a URL to a user profile. Furthermore, these data can be associated to more than one type of a record. For example, Pages, Posts and Comments can all be associated to a specific author. Yet, Author is not being considered as a separate record, as the search results are more likely to require the content published by the authors. Hence, it makes sense to separate the description of the Author from the tables describing the records. The same argument can be held for the published content. To make sure that searching through various types of information integrated into the published content can be organised, the content is being categorised, yet associated with all the relevant records – that are Posts, Pages and Comments.

3.2.7 Components of the Data Model

The requirements of the BlogForever project as captured in the agreement indicate that capturing the data associated with the core of the blog may not suffice. The inquiry into the blogs conducted as part of the development suggested, that the blogs represent a reach source of data. Hence, the core blog model was extended to be able to accommodate the data exhibited on blogs.

This section outlines the extended data model. It introduces additional entities that are grouped according to their nature. These groups are referred here as categories. The categories capture various aspects of blogs and provide a descriptive foundation to enable preservation of additional blog data. While the changes within the defined components are possible, they represent a necessary foundation that can be used for capturing additional information if necessary. An example for possible extension can be the integration of additional technical metadata fields into the Categorised Content for addressing the requirements of the project.

The categories enable storing the following types of blog data:

- Blog Context: descriptive data provided by the bloggers themselves.
- Network and Linked Data: a range of network data
- Community: information about the user base
- Categorised Content: descriptive data about the captured content
- Standard and Ontology Mapping: additional structures enabling mapping into other standards
- Semantics: information generated based on the analysis of the captured content.
- Spam Detection: spam mark-up and associated descriptive data
- Crawling Info: specifics about the crawling
- Ranking, Category and Similarity: various measures based on the analysis of existing data.
- Feed: information about the web feeds used

The graphical representation of the categories in relation to the blog core is presented in Figure 3.2-2.

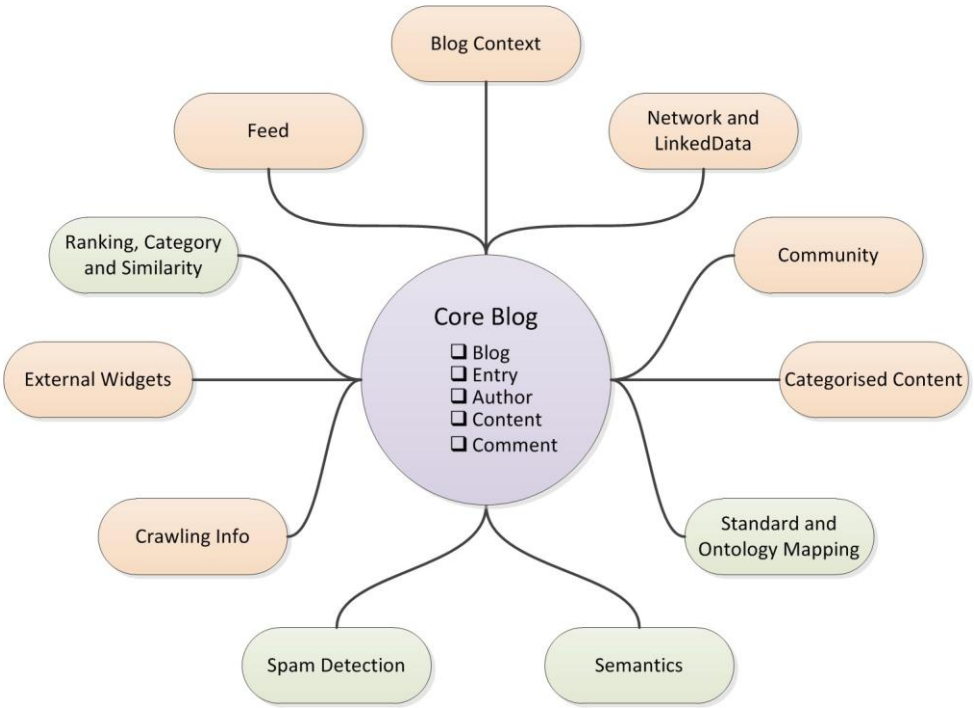


Figure 3.2-2: Blog Core and its Components

There are primary two types of categories described in the data model. The first type of category requires the data to be collected and extracted from the blog. The second type includes primarily derived properties and relies on the data already collected and stored in the repository. These two types of categories are represented in the diagram in different colours. The details about the each of the component are accessible from the original report (Stepanyan, 2011).

3.2.8 Representation in XML

For the purposes of the BlogForever project the proposed data model was represented using an entity-based approach. The entity-based approaches rely on the notion of an entity that represents an object in the real world. Information about the object is usually recorded as consisting of descriptive properties and relationships with other entities. Although entity-based approaches require a unique ontological view of the reality, these approaches are widely adopted and most frequently used. The tools and technologies that support entity-based modelling are also well established and accepted.

This notation of entity based approaches is widely adopted and is considered readable by both technologist and wider audiences. This representation demonstrates the primary entities as well as their attributes. While the primary and foreign keys have been collapsed to optimise the use of the limited graphical area, the relationships between the entities are demonstrated.

For the purposes of BlogForever, the data collected by the spider and stored in the repository might require transfer or delivery across a network. Among the approaches relevant for transferring data across software applications and networks is XML. XML is a widely adopted machine/human-readable mark-up language. Its simplicity and suitability for transactions over the network spawned a large number of XML-based languages and standards. Among those standards are METS[89] and MARCXML[90]. Both of the standards have been developed by the Library of Congress[91] and are widely for representing or describing the data.

However, the basic representation of XML data can be described as a tree, while the entity-based approach adopted for the development of the data model represents a graph-like structure. The follow up work from the above described model was to convert it to a tree-like XML structure that contains the entities and the properties. The converted XML files are presented in Appendix E. In addition to the blog data BlogForever also requires administrative data. The combination of the administrative and blog data represent the METS profile that has been discussed in Chapter 5. The use of METS is expected to contribute to the interoperability of the repository (as discussed in D2.3 (Kalb, Hendrik et al 2012) and further elaborated in the upcoming D3.2.

3.3 Significant Properties of Blogs: Bringing Together the Data Model and User Requirements

The quest for ‘significant properties’ is a common challenge arising within the digital preservation community. While the methodological frameworks for selecting these properties provide a good foundation, a continued discussion is necessary for further clarifying and improving the available methods. The discussion presented here is an effort to use the user requirement studies conducted in the project (Kalb et al. 2011) with the BlogForever data model (Stepanyan et al. 2011; also see Section 3.2) to identify the essential aspects of blogs that might need preservation support.

3.3.1 Disambiguation

In this deliverable we use the term "significant properties" to apply to two broad categories of digital content. The first use describes a view from the macro level (blogs as complex objects), and, the second, a view from the micro level (individual digital objects that occur in blogs).

In Section 3.3, the term is understood to apply to *an entire blog and its posts*. The study here is closely aligned with the Data Model and the survey of user requirements. It is intended to determine and identify the essential aspects of blogs that might need preservation support. Blogs are clearly complex objects. Significant Properties therefore can refer to a wide range of behaviours and performances of digital object types, the dependencies between these objects, links, structure, and so forth; in short all features which determine the complexity of blogs. Section 3.3 synthesises and consolidates this complexity to consider a minimum set of characteristics which must be preserved. Broadly, the properties here can be understood as *semantic and descriptive* terms, which we anticipate surviving in the preserved object and being presented through the BlogForever database.

In Section 3.4, the term "significant properties" is understood to refer to *individual digital object types* which might be found in blogs. The significant properties are understood to be those

[89] http://www.loc.gov/standards/mets/
[90] http://www.loc.gov/standards/marcxml/
[91] http://www.loc.gov/

properties of a digital object without which it would be unable to perform, or be rendered by an application or other process. This thinking is in line with the studies undertaken by the InSPECT project, which in turn was based on the NAA performance model. According to InSPECT, "an institution with curatorial responsibility for digital objects cannot assert or demonstrate the continued authenticity of those objects over time, or across transformation processes, unless it can identify, measure, and declare the specific properties on which that authenticity depends." In the case of image files, for example, there are seven core properties which must be identified, measured and declared. Rather than semantic, the properties in this instance are all entirely *technical* in nature. Five digital object types which occur commonly in blogs are identified in this section, and an overview of their significant properties is presented, based on work that has already been done by others in the field. We would expect to extend this micro model in later iterations, to include the significant properties of other (and more complex) digital objects that occur in blogs, such as 3D images, or even attached database files.

3.3.2 Related Work

The usage of the term "blog" within current discussions of social media often seems to suggest the existence of a coherent understanding of the term within the community. With the increasing number of blog-like services that encourage the propagation of user-generated content, the notion of a blog is becoming increasingly blurred (Garden, 2011). However, developing an understanding of a blog as an information object is invaluable, especially within the context of preservation initiatives that aim to capture the authenticity, integrity and usability of blogs.

This section positions the conducted study within the context of blog preservation by highlighting the limitations of the current practices and emphasizing the rationale for developing blog preservation solutions. It demonstrates the pressing need to identify the properties of blogs that need to be preserved prior to embarking on a task of preservation. The section proceeds to highlight the limitations within existing research on identifying these properties and proposes improvements accordingly. The section concludes by demonstrating the application of the modified approach on a use case and discussing the benefits and limitations of the proposed approach.

Hank and her colleagues (Sheble, 2007; and Hank, 2009) stress a range of issues that may affect blog preservation practices. The primary challenges of blog preservation are bound to the diversity of form that blogs can take and the complexity they may exhibit. A brief review of the literature shows that the definitions of blogs vary widely. The Oxford English Dictionary definitions of the terms ‘blog’ and ‘blogging’ highlight the temporal nature and periodic activity on blogs. Focus on technical elements of blogs is evident in the works by Nardi and his colleagues (Nardi 2004, p. 43). Other definitions, for instance by Pluempavarn and Panteli (2008, p. 200), deviate from a standpoint that looks into the technical aspects of blogs and into the socio-cultural role of blogs. The capacity of blogs for generating social spaces for interaction and self-expression (Lomborg, 2009) is another characteristic. The social element of blogs entails the existence of networks and communities embedded into the content generated by bloggers and their readership.

Due to the complexity of the Blogosphere - as shaped by the variety of blog types, the changing nature of blog software and Web standards, and the dependency on third party platforms - it is likely that lossless preservation of blogs in their entirety is unrealistic and unsustainable. Blog preservation initiatives should, therefore, question what essential properties they must retain to avoid losing their potential value as information objects. It becomes eminent that gaining insight into the properties of blogs and their users is necessary for designing and implementing blog preservation systems. The quality of the preserved blog archives is dependent on capturing the fundamental properties of blogs. The following question would then be: what methods should be used for identifying these properties?

3.3.3 Significant Properties: an Attempt to Measure Preservation Performance

In the digital preservation community, one of the prevailing approaches for defining what to preserve is bound to the notion of significant properties[92] (see also Hedstrom and Lee, 2002). It is argued (Deken 2004) that significant properties can help define the object and specify what to preserve, before deciding how to preserve. It has been acknowledged (Knight and Pennock, 2005), however, that defining the significant properties without ambiguity remains difficult. The main problem is the lack of a suitable methodology for identifying the significant properties. While there are tools and frameworks for defining and recording technical characteristics of an object, Tyan Low (2011) argues that identifying significant properties in general still remains contended, primarily due to the methods employed for the task. Low *(ibid.)* outlines the list of projects that attempted to develop mechanisms for identifying significant properties. The outcomes of these projects led to a range of frameworks and methodological tools, such as PLANETS[93] Plato that focuses on stakeholder requirements (Becker 2008), InSPECT that combines object and stakeholder analysis (Knight and Pennock 2009), a JISC[94]-funded initiative that continues the discussion (Hockx-Yu and Knight, 2008), and a template of characteristics (NARA 2009) developed by NARA[95].

Yet, despite the seemingly large number of tools that exist for organising significant properties into a range of types, expressing them formally, and testing their fidelity when subjected to selected operations (such as migration and emulation), the approaches available for guiding the decision making processes in identifying the relevant types and properties remain too abstract, especially with respect to complex objects (Farquhar, 2007).

However, considering the range of available solutions, InSPECT framework (Knight and Pennock 2009) is considered to offer a more balanced approach to identifying significant properties (Tyan Low 2011). The advantage of this approach is encapsulated in the parallel processes it offers for analysing both the object and the stakeholder requirements. The framework is claimed to support identification of the significant properties of information objects by progressing through a specified workflow.

The InSPECT framework stands out as one of the first initiatives to accentuate the role of object functions derived from an analysis of stakeholder requirements as a gateway to identifying significant properties of digital objects.

InSPECT (Knight and Pennock 2009) is built on the Function-Behaviour-Structure framework (FBS) (Gero 1990) developed to assist the creation and redesign of artefacts by engineers and termed useful for identifying functions that have been defined by creators of digital objects. The workflow of InSPECT is composed of three streams: Object Analysis, Stakeholder Analysis, and Reformulation. Each of these streams is further divided into stages that are argued by the authors *(ibid.)* to constitute the process of deriving significant properties of a preservation object.

However, the InSPECT framework was originally developed in line with simple objects such as raster images, audio recordings, structured text and e-mail. The main limitation of the framework, as discussed by Sacchi and McDonough (2012), is its reduced applicability for working with complex objects. They *(ibid., p. 572)* argue that the framework lacks "the level of granularity needed to analyze digital artifacts that — as single complex entities — express complex content and manifest complex interactive behaviours". Similar complexities exist in the context of blogs,

[92] http://www.leeds.ac.uk/cedars/
[93] http://www.planets-project.eu/
[94] www.jisc.ac.uk/
[95] http://www.archives.gov/

making application of InSPECT in its current form challenging. Hence, we propose a set of adjustments into the framework to improve its capability of working with objects like blogs.

The Object and Stakeholder Analysis are considered to be the two parallel streams termed as Requirements Analysis. Each of the streams results in a set of functions that are cross-matched later as part of the Reformulation stage. To address the limitation of InSPECT, we first focus on the lack of detailed instructions for conducting Object Analysis. The framework suggests the possible use of characterisation tools or technical specifications for the purpose of object structure analysis (Section 3.1 of (Knight and Pennock 2009). These suggestions presuppose the existence of such a tool or specification. While such a tool or specification may be available for fairly simple self-contained digital objects, like electronic mail, raster images, digital audio recordings, presentational mark-up, the situation is less straightforward for complex digital objects, such as weblogs and/or other social network media. In addition to the lack of guidance in defining the object structure, the framework suggests identifying functions associated with object behaviour as part of the object analysis. These functions are then proposed to be consolidated with those identified from the stakeholder analysis stream. Consideration of functions introduces an ambiguously defined stakeholder view as part of the object analysis. This ambiguity and a higher level of abstraction when working with functions leads us to propose modifications of the framework to enable its application in the context of blog preservation.

3.3.4 Proposed Changes

The modifications discussed here, firstly, introduce an ontological perspective into the Object Analysis stream and, consequently, further clarify the degree of overlap between the two streams of analysis. Secondly, it proposes integrating results from two separate streams at the level of properties rather than functions. We elaborate the proposed changes further down in this section. We justify the changes introduced into the Object Analysis stream and clarify the subsequent adjustments to the workflow of the framework in the remaining part of this section. We then demonstrate the application of the framework by presenting a use case on blogs and discuss our experience in employing this approach.

The modifications introduced in the Object Analysis stream aim to address the limitation of InSPECT (i.e. base framework) in specifying appropriate procedures for performing the analysis of complex objects and identification of their properties. We propose adopting an ontological perspective, to eliminate the impediment of the framework for guiding the preservation of objects such as blogs. Unlike simpler objects of preservation, such as images or text documents, blogs are usually comprised of other objects or embedded elements and demand a more structured approach when analysing these to avoid overlooking important properties.

The use of ontological perspectives is common in data modelling and has recently been receiving attention in the area of digital preservation. For instance, Doerr and Tzitzikas refer to a set of ontologies, such as DOLCE, OIO and CIDOC CRM, established and commonly used in (digital) libraries, archives and related research initiatives. They *(ibid.)* argue that the use of ontologies makes the process of understanding sensory impressions of information objects more objective. Indeed, an ontological perspective can enhance the process of object analysis by offering abstraction to the level of conceptual objects along with the formalism for describing the structures of the compound objects. In contrast to current digital preservation research, Doerr and Tzitzikas *(ibid.)* emphasise the possible range of information objects (and relevant features) encompassed within a single information carrier and argue for exploring the sensory impressions rather than the binary forms objects. However, stakeholder views are not directly discussed in the work by Doerr and Tzitzikas (ibid.). We attempt to follow Doerr's suggestion and integrate it with InSPECT. This enables us to use an ontological perspective for exploring complex objects (i.e. identifying compound objects and relationships among them) in addition to conducting a stakeholder analysis. The two streams of analysis can then be consolidated to inform the preservation strategy.

The diagrammatic representation of the proposed framework is presented in Fig. 3.3-1. The workflow of the framework initiates with the selection of the object of preservation and proceeds, via two parallel streams, to guide the Object and Stakeholder Analysis. The Object Analysis aims to establish the technical composition and the structure of the preservation object. This stage starts with the analysis of object structure. It focuses on the essence of object of preservation and aims to identifying both conceptual and logical components of this compound object (viewed as classes). The next stage focuses on identifying relationships between the identified components. The relationships that exist between object components are expected to be explored and documented at this stage. Once the components and the relationships between those are identified, the properties of the object can be elicited and documented. The properties of the objects of preservation have to capture the characteristics of the compound objects along with their technical specifications. The stream of Object Analysis is therefore expected to result in developing a set of documented compound objects and associated properties that are to be cross-matched and refined with the outcomes of the parallel stakeholder analysis stream.

The Stakeholder Analysis aims at identifying a set of functions that stakeholders may be interested in and, subsequently, derive the properties of the preservation object that would be necessary to capture for supporting the required functions. The analysis starts with the identification of stakeholder types. They can be discovered through the investigation of policies, legal documents or communities related to the object. This stage is followed by the contextualisation of the object, which highlights stakeholders' perceived differences or variations in the levels of object's granularity. The third stage aims to determine the behaviour, which can be accomplished by examining the actions taking place in the real world. Having identified the actual behaviour, the anticipated behaviour is recorded through a set of functions. The last stage of the stakeholder analysis enables eliciting the properties of the object that are essential for satisfying the stakeholder requirements. The following stage aims at assessing and cross matching the properties identified from the two parallel streams of Object and Stakeholder Analysis.

The process of Cross-Matching and Refinement enables the consolidation of the identified properties and their refinement into an extended list of properties. The consolidation of the two independent streams is proposed to be conducted at the level of properties (rather than functions) and aims at integration of identified properties. The refinement of the integrated list of properties leads to the proposal of properties to be considered for preservation. As significance is (repeatedly) associated with stakeholder views (Dapper and Farquhar 2009) the outcomes of the stakeholder analysis should remain in constant focus. The refinement of the integrated list should prioritise the inclusion of properties identified from the Stakeholder Analysis stream.

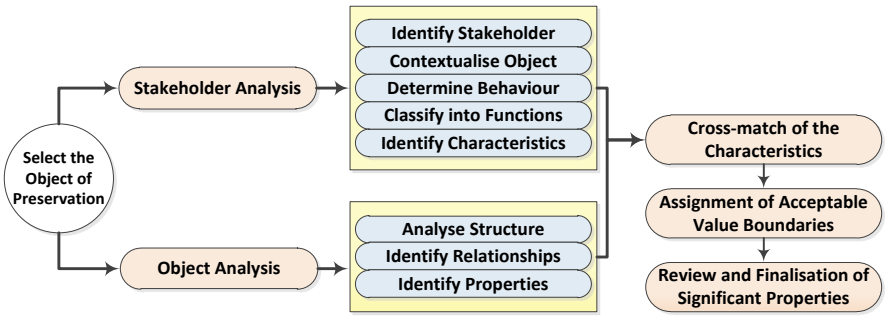


Fig. 3.3-1: Modified version of the base framework.

The Review and Finalisation stage includes the reflection on the previous steps and consideration whether any revisions are necessary. At this stage, identified properties can be recorded and the boundaries of their values can be assigned. The properties can then be used to define the objects of

preservation and to progress with the design and development of the preservation initiative (for instance, for developing the databases necessary for storing data).

3.3.5 Applying the Proposal to Blogs

The rationale for the identification of significant properties may lie in "ensuring the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record" (Knight and Pennock 2009, Wilson 2005). However, its outcome directly feeds design, development and planning of preservation solutions. This section integrates and consolidates some of the work carried forward as part of a blog preservation initiative (Stepanyan 2011, Kalb 2011). It describes the process of Object Analysis conducted to explore the object of preservation and (in the subsequent section) Stakeholder Analysis from the interviews exploring anticipated functionality of a blog repository.

Object Analysis

Blogs exhibit a considerable diversity in their layout, writing style or organisation. The analysis of this complex object, therefore, can be conducted from various perspectives and at different levels. Object analysis can employ an approach, widely accepted within the preservation community, that describes an information object as a conceptual (e.g., as it is recognised and perceived by a person), logical (e.g., as it is understood and processed by software), and as a physical object (e.g., as a bit stream encoded on some physical medium) (Thibodeau 2002). In this section we present our approach adopted for the case of blogs and discuss this experience in a broader context. Identification of generic concepts of an object, their compound structures, hierarchy and relationships (without necessarily reflecting the operations expected to be performed) is common in ontology and data modelling. It can be used for the identification of generic concepts, subsequently leading towards the identification of object's properties (Dillon et al 2008). A structured and iterative approach was adopted, to review and refine the analysis of the blog object. An alternative to this approach would involve consideration of an existing ontology. In this case, we conducted the following: [a] an inquiry into the database structure of open source blog systems; [b] an online user survey (900 respondents) to identify important aspects and types of blog data in the current usage behaviour; [c] suggestions derived from recent developments and prospects for analysing networks and dynamics of blogs; [d] an inquiry into the technologies, formats and standards used within the Blogosphere; [e] an inquiry into blog structure based on evaluation of blog feeds (2,695 in total); and [f] an inquiry into blog APIs.

As a result of the above mentioned inquiries, a coherent view on the concepts of the blog object was acquired, informing further development of a respective data model. It enabled understanding the structure of blogs and help identifying their components, relationships and properties. The rest of this section outlines the process of conducting object analysis. Given the space limitation, a complete account of the performed study is omitted here. We briefly outline the conducted work, the details of which are available elsewhere (see Stepanyan 2011).

Database Structure, User Views and Network Analysis

The knowledge of the domain, user survey and inquiry into conceptual models of blogs and their networks enabled identifying the most prominent conceptual and logical objects. Blogs may contain Entries (identified as being either Posts or Pages) that may have Comments and are associated with an Author. Both Entries as well as Comments exhibit certain Content. These entries are analysed further and (where relevant) broken down into smaller compound objects. For instance Content, as one of the most complex elements is further described by simpler objects like Tags, Links, Text, Multimedia, etc. For demonstration purposes, we use only most frequently occurring components

that are: Entry (Post/Page), Comment, Content, Author and the Blog itself, omitting the details due to space constraints.

In addition to the identification of compound entities of the complex objects, it is necessary to study the relationships that exist across these entities. This is particularly relevant when working with blogs, which are known to become interaction platforms and weave into complex networks. The structural elements of blogs, as conceptual, logical or physical objects, can represent the nodes and attributes, or define the ties of various networks. An insight into the network structures within and across blogs can be important gaining insight into the conceptual and logical objects. Identification of properties that may be of interest to archivists can greatly benefit from an insight into the network aspects of blogs and their components.

For instance, identifying different ways of citations within blogs can provide insight into the inter-related structure of objects, such as entries, comments or authors. However, while links added across blog posts may be technically similar to those added via link-back mechanisms, the ties formed by these two different types of links may be approached or treated differently. Our experience with this use case infers that the analysis of a blog in relation to others provides information about the properties of blogs and becomes useful as part of the Object Analysis stream. Furthermore, the theoretical and technological advances of analysing blogs and their networks should also be considered for gaining insight into the blogs and the phenomenon of blogging in general.

Technologies, Formats, RSS Structure and APIs

While identification of compound elements and understanding of their relationships is an important step, it constitutes a high level view. To continue the analysis of the object and identify potential properties for preservation, a lower level view on the conceptual and logical objects is necessary. An inquiry into technical aspects of blogs provides information about the lower level properties of the affiliated objects. To demonstrate this in the context of this use case, we highlight some examples of eliciting the properties of the blogs components.

To discuss an example of lower level properties we could consider the textual content. Textual content can be represented as a set of characters, along with its stylistic representation (e.g. font, colour, size), encoding, language, and bit stream expressed on the selected medium. The lower level description primarily deals with files, and can inform their storage and retrieval. Therefore, analysing the HTML code of blogs can reveal details about the technological backbone of blogs (formats, technologies, standards), which remains invisible to most blog users. Empirical studies exploring the physical objects can be particularly helpful in identifying potential properties. We briefly outline an example of a study to demonstrate the relevance of this approach.

An evaluation of 209,830 blog pages has been performed and reported (Banos 2012). The HTML-based representation of these resources was parsed and searched for specific mark-up used to define character sets, languages, metadata, multimedia formats, third-party services and libraries. The quantitative analysis of certain properties exhibited by the specific objects allowed us to describe common properties exhibited in blogs within the Blogosphere.

The evaluation was particularly useful in identifying properties of various compound objects (e.g. Content, which was further broken down into smaller logical objects and respective characteristics of associated physical ones). Geographical location (GPS positioning), as a contextual characteristic associated to Blog Entries or Content, was another direct outcome that emerged from the above evaluation. For instance, properties identified for the object Entry, and used in for demonstration purposes in this use case, include: [a] Title of the entry; [b] Subtitle of the entry; [c] Entry URI; [d] Date added; [e] Date modified; [f] Published geographic positioning data; [g] Information about

access restrictions of the post; [h] Has a comment; [i] Last comment date; and [j] Number of comments. A more detailed description of the conducted analysis, as well as the complete list of objects and properties is made available elsewhere (Stepanyan 2011) due to space constraints.). The properties (excluding those associated with omitted objects), identified as part of the Object Analysis phase are presented in Table 3.3-1.

Table 3.3-1: Most common blog objects and their characteristics

Object	Significant Characteristics
Blog	Title of the blog
	Subtitles of the blog
	URI of the blog
	Retrieved language field, as defined by the blog
	Retrieved charset field, as defined by the blog
	Platform of the blog powering service, retrieved where available
	Versioning information about the platform
	IP address of the blog
	Hosting location: city and country details
	Last activity date as retrieved from the blog
	Post frequency as retrieved from the blog
	Update frequency as retrieved from the blog
	Notes of copyright, ownership, distribution and access rights.
Entry	Title of the entry
	Subtitle of the entry if available
	Entry URI
	Date added
	Date modified
	Geographic positioning information
	Information about accessibility of the post
	Has a comment
	Last comment date
	Number of comments
Page	Design template (if available and if different) from the general blog
Post	Type of the post if specified: attachment or other custom type
	Information about the service used for posting if specified
	URI to the previous post if available
	URI to the next post if available
Comment	Subject of the comment as retrieved
	URI of the comment if available
	Date comment was added
	Date comment was modified
	Geographic positioning information
	Has a comment
	Number of comments
	Is child of the parent post
	Is child of parent comment
Au-thor	Author name as displayed
	Author email address as displayed
	Is anonymous: boolean property
Cont ent	Content as extracted
	Content format (i.e. HTML, XML)

	Additional notes if available
	Information on encoding of the content
	Notes of copyright, ownership, distribution and access rights.

Stakeholder Analysis

The objective of the Stakeholder Analysis is to identify a set of functions that stakeholders may be interested in and, subsequently, derive the properties of the preservation object that would be necessary to capture for supporting the required functions. The initial task was to identify or acknowledge the stakeholders that may interact with an instance of the object of preservation or their collection as part of a repository. Stakeholder interviews for identifying their requirements are an essential part of Stakeholder Analysis. Their methodological foundations as well as the complete list of functional requirements are available in the BlogForever deliverable D4.1 User Requirements and Platform Specifications Report (Kalb 2011). A brief outline of the process directly affecting this use case is presented below.

Identification of Stakeholders

Within the context of blog preservation we acknowledge three groups of stakeholders: Content Providers, Content Retrievers and Repository Administrators. Within each of these groups we identified individual stakeholders: [a] Individual Blog Authors; [b] Organizations within the Content Providers group; [c] Individual Blog Readers; [d] Libraries, Businesses; [e] Researchers within the Content Retrievers group; and finally, [f] Blog Hosts/Providers and [g] Organizations (as libraries and businesses) within the Repository Administration group. This extensive list of stakeholders can be justified by the multitude of ways (including some unknown ways) of using preserved objects by present and future users (Yeo 2010). Hence, rather than selecting a single definitive solution, it remains important to identify a range of essential as well as potential requirements to maximize the future usability of a blog repository. A user requirement analysis was performed for every stakeholder type. It focused on analysing stakeholder interaction with blogs via digital repository software.

Applied Method of Requirement Analysis

There is a range of methods for conducting effective user requirement analysis (Hull 2010). In the context of this study we conducted an exploratory, qualitative study by means of semi-structured interviews. A set of stakeholders, from each of the groups, was approached to be interviewed. The structure of the interviews was designed to enable consolidation of the results across the stakeholders and stakeholder groups. General methodological and ethical guidelines for conducting qualitative inquiry of this kind were followed.

A total of 26 interviews were conducted. Candidate interviewees were identified and approached individually. The sample of interviews was selected in a way that each of the defined stakeholder groups was represented by at least one interviewee. The distribution of interviewees for each of the stakeholder groups was: 10 for Content Providers; 12 for Content Retrievers; and 4 for Repository Administrators. The requirements were then analysed and a set of user requirements was identified.

Identified Requirements and Properties

The analysis followed a three-step approach. Initially, each interview was analysed regarding the indication of requirements in the two main categories functional and non-functional. The non-functional requirements were classified into: user interface, interoperability, performance, operational, security and legal requirements. Subsequently, the requirements were analysed for recurrent patterns, aggregated and further clarified. The final list of identified requirements included a list of 115. Further details discussing the methods and the complete list of elicited requirements is available elsewhere (Kalb 2011). The requirements that depend on existence of certain data elements were then shortlisted as shown in Table 3.3-2.

Table 3.3-2: A sample list of requirement functions identified from stakeholder interviews. ***(*FR: Functional Requirement, EI: Interface Requirements, UI: User Requirements, RA: Reliability and Availability Requirement)***

Req.	Description	Req. Type*
R12	Unique URI with metadata for referencing/citing	FR/UI
R17	Distinguish institutional/corporate blogs from personal blogs	FR
R18	Display blog comments from several sources	FR
R19	Display and export links between/across blog content	EI/UI
R20	Prominent presentation of citations	FR/UI
R22	Historical/Chronological view on a blog	UI

Identifying data elements that are necessary for the implementation of the requirements leads to properties of the preservation object that can be attributed as important. Hence, the requirement analysis, in this case, proceeded in identifying data elements and conceptual entities they are associated with. The identified data elements are presented in Table 3.3-3. The properties elicited from the Stakeholder Analysis were then cross-matched with those resulting from Object Analysis stream and further refined into a consolidated list of properties.

Table 3.3-3: Properties elicited from stakeholder requirements

Req.	Objects	Identified Properties
R12, R20	Entry	Digital Object Identifier(DOI)/Unique Identifier(UI)
R17	Blog	Blog type
R18	Comment	Comment type, source URI, service name
R19	Content	URI, URI type (e.g. external/internal)
R22	Blog, Entry, Comment	Creation/Capture/Update dates and time, time zone, date/time format.

Bringing Together Object Analysis and Stakeholder Analysis

The next step towards consolidating the list of properties includes the process of cross-matching, integration and refinement. The properties, identified from the two streams of Object and Stakeholder analysis are being compared and integrated into a single set of properties. It requires cross-matching and integration of properties that were missing from either of the list and eliminating same properties that were listed with different names.

We bring an example of cross-matching by referring to the property of DOI/UI96 for an entry, which has been identified from Stakeholder Analysis, but did not surface in Object Analysis. Unlike URIs that also constitute a unique identifier, an alternative approach similar to DOI was identified as necessary from the Stakeholder Analysis. Offering a consistent approach to referencing that is detached from the original resource differentiates between these identifiers. Hence, DOI/UI constitutes a property that is necessary for developing a system that meets stakeholder requirements. As a result, the property is added to the integrated list. This example demonstrates that Stakeholder Analysis allowed complementing the Object Analysis stream, which remained confined to intrinsic attributes of an entry such as URI.

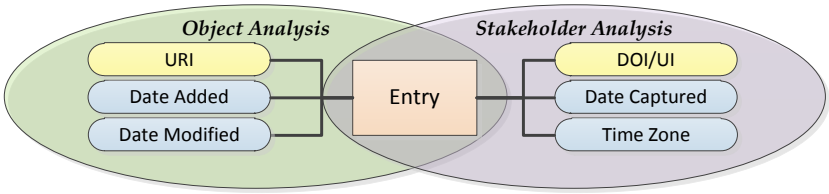


Fig. 3.3-2: An example of cross-matching and integration of properties, which were identified from the two parallel streams of Object and Stakeholder Analysis.

The requirement for providing a historical/chronological view of the entries, demonstrates another example where in addition to having the date and time of publication/editing, information about the time zone and date of capture is shown to be important. This can be elicited from the requirement R22 as shown in Table 3.3-3. While dates have already been identified from the object analysis, their alignment within the repository that takes into account the time zone differences has been identified as important from the stakeholder analysis. The examples of cross-matching and integration are illustrated in Fig. 3.3-2.

The final stage of the framework suggests to review the information collected at the previous stages and to decide whether additional analysis is necessary. The process of the review can be considerably improved if acceptable value boundaries are assigned to the identified properties. For instance, in line with the previous example, acceptable values and recognized standards can be considered for capturing the Time Zone and Date. Reflecting on acceptable boundaries can attest to the need for breaking down compound properties or reviewing the properties before their finalisation. The less abstract the identified properties are, the easier it would be to progress to the implementation of the preservation initiative. Returning to the Stakeholder Analysis and shortlisted requirements can reaffirm the identified properties or lead to further extension.

3.3.6 Discussion and Conclusions

The use case (Section 3.3.5) represents an example of applying a methodological framework and informing a blog preservation initiative. It enables us to advance the discussion on identifying significant properties of complex objects such as blogs. Reflecting on our experience of the process of identifying and consolidating the object properties we report the benefits and disadvantages of employing this framework and suggest directions for further research.

The integration of the ontological perspective into the Object Analysis stream of the framework has indeed enabled a thorough analysis of the compound object under study. The results of object analysis produced a fine grained representation of the compound blog object. Integration of the ontological perspective into the InSPECT framework provided the lacking methodological guidance for working with complex objects. Furthermore, the modification of the framework that enabled cross-matching Object and Stakeholder Analysis streams at a lower level of properties has also been

[96] http://www.doi.org/

demonstrated beneficial. It clarified the process of comparison due to the use of specific properties rather than more abstract (higher level) functions.

However, the modified approach still lacks unambiguous methodological guidance for defining significance associated with each of the identified property. Supporting the identification of properties that are not significant will also be a useful addition to the framework. Potential directions for future work may involve developing tools for guiding stakeholder analysis and defining the levels of significance associated with properties. Exploring the possibilities of discussing the concept of significance as a relative spectrum should also be followed as part of the future research.

This section advanced the discussion on the topic of significant properties that engages the preservation community. It positioned the conducted inquiry within the context of blog preservation. Highlighting the limitations of current approaches in preserving blogs, this section defined the rationale for understanding and defining blogs as objects of preservation.

Building on the body of work that provides methodological foundations for identifying significant properties, this section adapted the recently developed InSPECT framework (Knight 2009) for enabling its use with complex objects. It proposed to employ an ontological perspective on analysing compound objects enabling systematic analysis and de-composition of blogs into components and understanding the relations between them. This approach was demonstrated to be beneficial, leading towards identification of compound entities and properties of blogs. The modifications provided further clarification into the streams of Object and Stakeholder Analysis. Instead of cross-matching the functions, the framework proposes to consolidate the results at a lower and more tangible level of properties.

While the use case demonstrated the applicability of the modified framework on the complex blog objects, it also highlighted a number of limitations. More specifically, further clarification is necessary for identifying properties that should not be considered for preservation. The development of methodological tools for defining and measuring significance is particularly important. Future work can also extend the discussion on automating the process of identifying these properties. The reuse and integration of existing ontologies is another direction that requires further examination. Nevertheless, the results discussed here support the argument that the proposed modifications enhance the base framework by enabling its use with complex objects, and provide insight for advancing the discussion on developing solutions for identifying significant properties of preservation objects.

3.4 Significant Properties of Embedded Digital Object Types

The BlogForever weblog survey (Section 3.1) and the BlogForever data model (Section 3.2) identified several *digital object types* and *categorised content* that figured prominently within weblogs. In this section, we present significant properties of these objects determined by previous research. The discussion here is limited to a description of "*the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record.*"[97] We will not delve into a discussion of tools developed to extract information required to quantify these properties. This topic will be revisited in Chapter 6.

The discussion of significant properties stretches back considerably but one of the first formal definitions may be traced back to the Cedars Project[98] who defined it as "those technical characteristics agreed by the archive or by the collection manager to be most important for

[97] http://www.dpconline.org/docs/events/080407sigpropsWilson.pdf
[98] http://www.ukoln.ac.uk/services/elib/projects/cedars/

preserving the digital object over time"[99]. Quantifying these properties to measure the effectiveness of a selected preservation strategy[100] was adopted as a component in the workflow developed by the PLANETS project[101] and the subsequent development of the PLATO preservation planning tool[102]. Discussions about the subjective nature of these properties (e.g. Dappert & Farquhar 2009) are also noteworthy. The abundance of discussion on these properties has led to its inclusion in the reference model for OAIS[103] as Transformational Information Property, "whose preservation is regarded as being necessary but not sufficient to verify that any Non-Reversible Transformation has adequately reserved information content", requiring further representation information including semantic information[104].

Investigating the Significant Properties of Electronic Content over Time (InSPECT) aimed to expand and articulate the concept of 'significant properties', determine sets of significant properties for a specified group of digital object types, evaluate methods for measuring these properties for a sample of relevant representation formats, investigate and test the mapping and comparison of these properties between different representation formats, and identify any issues requiring further research. Significant Properties are defined by InSPECT as *The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record.* The definition was tested by identifying Significant Properties (SPs) in particular digital object types. The project chose audio, email, raster image and structured text objects as distinct classes of digital objects, and typically three formats for each type. After developing a framework for considering SPs, they were determined for each object type using a consistent methodology and extracted from sample files. Transformations (in this case through file format migration) were undertaken and SPs were extracted from the resultant objects. Comparisons were then made of the source and target objects to see how well the target retained the significant properties of the source.

The methodology used by this project is explained in their report. They worked to the Wilson metaphor of "performance". They undertook research of literature including Rothenberg & Bikson (1999), the CEDARS Project, the CAMiLEON project, the National Archives of Australia, RLG, Digital Preservation Testbed, DELOS, as well as more recent developments by the CASPAR, PLANETS and four JISC-funded Significant Properties projects. The conceptual Utility Analysis and Objective Tree (Rauch, Strodl & Rauber, 2005) was developed and applied in the DELOS research and refined in the PLANETS projects as a metric to test and evaluate digital preservation strategies.

In the following we summarise the properties identified as a culmination of these past initiatives for five of the six digital object types defined in Section 3.1: structured text, image, document, audio, and moving image. Work is still on-going to determine significant properties with respect to scripts and executables.

3.4.1 Structured Text

37 properties were identified by the InSPECT project. Assessment of the significant properties of structured text was based primarily on the latest W3C HTML 4.01 specification. Many elements were considered significant "in certain circumstances". Body colour text (Text=[colour]) illustrates the reasoning behind this decision: it is an attribute that specifies the foreground colour for text on the page. Web Accessibility Initiative guidelines deprecate the use of colour alone to convey

99 http://www.webarchive.org.uk/wayback/archive/20050410120000/http://www.leeds.ac.uk/cedars/guideto/collmanagement/index.html

100 http://www.springerlink.com/content/c20150x260758r11/

101 http://www.planets-project.eu/

102 http://www.planets-project.eu/

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

104 http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

information for accessibility reasons so it should not be considered significant. Some authors, however, use colour artistically and others choose to use it to convey semantic meaning, e.g. using red to indicate a negative number. These elements have not been included in the following list, but it may need to be augmented by some of them in specific circumstances. HTML 3.2, HTML 4.1 and XHTML 1.0 were selected as the formats for testing as these were all supported by the JHOVE tool which was chosen to do the file characterisation.

Table 3.4-1: significant properties of structured text

	Semantic Unit
1	Title
2	Creator
3	Date
4	Keywords
5	Rights
6	Div
7	Span
8	Language
9	Paragraph
10	Line break
11	Headings
12	Emphasis
13	Bold
14	Italics
15	Underline
16	Strong emphasis
17	Strikethrough
18	Horizontal Rule
19	Inserted text
20	Deleted text
21	Samp
22	Cite
23	Defined Terms (DFN)
24	Code
25	Abbreviation
26	Acronym
27	Quotations
28	Subscript / Superscript
29	Address
30	Button
31	List Elements
32	Table Elements
33	Image
34	Link
35	Applet
36	Frame
37	Frameset

Relevant W3C standards

- W3C HTML 4.01 specification
- W3C Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification
- W3C Extensible Markup Language (XML) 1.0 (Fifth Edition)
- XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition)

3.4.2 Image

Seven properties were identified by InSPECT. The ANSI/NISO Z39.87 data dictionary was used as the basis for the project team's analysis of the significant properties of raster images. TIFF (Version 6.0), JPEG (Version 1.02) and GIF (Version 89a) were the formats chosen for testing as these were all supported by the JHOVE tool chosen to perform file characterisation. They are also all widely used in different settings as raster image representation formats.

Table 3.4-2: significant properties of images

	Semantic Unit
1	Image Width
2	Image Height
3	X Sampling Frequency
4	Y Sampling Frequency
5	Bits per sample
6	Samples per pixel
7	Extra samples

See also

JISC Digital Preservation Programme: *Study on the Significant Properties of Vector Images* (Coyne, 2007)

Other relevant standards

- ANSI/NISO Z39.87 data dictionary

3.4.3 Document

The report *Document Metadata: document technical metadata for digital preservation* (Chou, 2009) proposes 12 significant properties for documents. Formats in scope included but were not limited to doc, pdf, odt, sxw, sdw, wpd and wps. For each metadata element listed in the data dictionary, the document formats are listed that are known to contain either the associated metadata values directly in the file or that could be determined indirectly by parsing the files.

Table 3.4-3: significant properties of documents

	Semantic Unit
1	PageCount
2	WordCount
3	CharacterCount
4	ParagraphCount
5	LineCount
6	TableCount
7	GraphicsCount
8	Language
9	Fonts
10	FontName
11	IsEmbedded
12	Features

Relevant ISO standards

- Open Office: ISO/IEC 26300:2006
- PDF: ISO 32000
- OOXML (Microsoft): ECMA-376, ISO/IEC 29500

3.4.4 Audio

14 properties were identified by InSPECT. 1-6 are the core properties; 7-14 only apply if the audio recording contains BEXT-formatted metadata. MPEG-1 Audio Layer 3 (MP3), Microsoft Waveform (.wav) and Broadcast Wave (BWF) were selected as audio representation formats for testing. Several specifications were considered in compiling the set of significant properties, including the draft AES-X098B specification, Harvard University Library's DRS administrative metadata for digital audio schema, PBCore and the Library of Congress AudioMD schema, as well as the preservation guidance provided by the Indiana University Digital Library Sound Directions project, Council on Library and Information Resources & Library of Congress, Arts and Humanities Data Service and CDP Digital Audio Working Group.

Table 3.4-4: significant properties of audio

	Semantic Unit
1	Duration
2	Bit depth
3	Sample rate
4	Number of channels
5	Sound field
6	Sound map location for each channel
7	Description
8	Originator
9	OriginatorReference
10	OriginationDate
11	OriginationTime
12	Coding History
13	Quality Report
14	Cue Sheet

3.4.5 Moving Image

15 properties were identified by the JISC Study on the Significant Properties of Moving Images (Coyne, Stapleton 2008) [105]. This study drew on the work of CEDARS, CAMILEON and InSPECT.

Table 3.4-5: significant properties of moving images

	Semantic Unit
1	imageStreams
2	audioStreams
3	Length
4	Width
5	Height

[105] http://www.jisc.ac.uk/media/documents/programmes/preservation/spmovimages_report.pdf

	Semantic Unit
6	bitDepth
7	colourModel
8	colourSpace
9	pixelAspectRatio
10	frameRate
11	Lossless
12	compressionRatio
13	Codec
14	Interlace
15	Metadata

3.5 Conclusion

In this chapter, we discussed four different approaches to examining weblogs (survey, data modelling, user requirements, and previous research results in object analysis). The investigation leads to

- An idea of the variety of objects and/or entities that we might need to support (Section 3.1).
- A data model of how the objects and/or entities are positioned in relation to each other. In particular, we are able to propose four repository record types for the BlogForever repository (Section 3.2).
- A set of weblog properties recommended for preservation in order to meet current user requirements and make the weblog meaningful (Section 3.3).
- A set of properties for selected digital object types from previous research initiatives that we recommend for preservation in order to support the correct rendering of objects.

.

The investigation in Section 3.3 and it relationship to Section 3.4, brings to light the complex interconnected structure of weblogs, and emphasises the necessity of a multi-level investigation of weblog properties on the macro level (components in the pages to be retained to meet user requirements) and micro level (to enable the correct render target embedded objects).

It is important to note, however, that the micro level object types can occur within many different parts of the macro level structure (e.g. post, comment, and/or page). That is, while each instance of the digital object type may be introduced to express one or more aspects of syntax, semantics, and/or pragmatics, the digital object type a priori is not tied to any of these. We will show evidence in the next chapter that, the way the two types of properties interact is defined by the notion of a *weblog community*, defined by technical characteristics, link network structure, and information being shared. In particular, by showing how technical characteristics are shared within communities, we will show how the technology itself forms a part of the cultural heritage we aim to preserve. Also, by discussing the varying complexities of weblogs across communities we will be able to better formulate preservation experiments in terms of representative data to address scalability and performance.

In this chapter we limited the discussion to characteristics of the weblog content itself. That is, we did not discuss the difficulties arising from the dependencies of the weblog content on software/hardware environments such as web browsers and operating systems. These issues will be raised in the next chapter.

4 Preservation Strategy Testing

The objective of preservation testing is to test whether the strategy that the repository has adopted provides *sustainable access* to repository holdings in a way that instantiations of the materials accepted into the repository are within an *acceptable distance* away from the *expected syntax, semantics, and pragmatics* associated to the object (cf. Ross 2006*).

A previously proposed generic workflow for preservation strategy testing is illustrated in Figure 3.5-1.

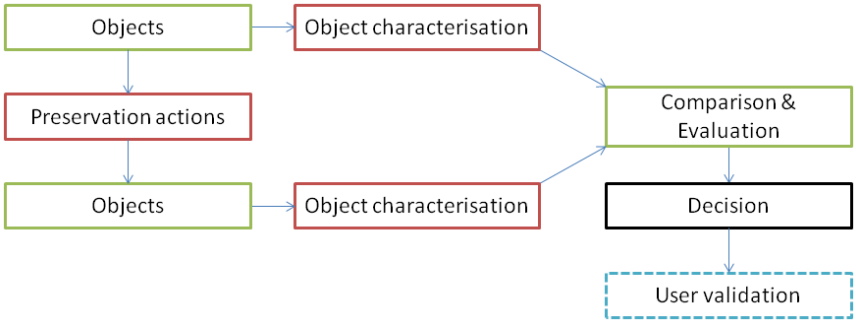


Figure 3.5-1 Generic workflow for digital preservation experiments

The workflow represents the bare-bone concepts derived from a combination of several different works (Strodl et al. 2006; Aitken et al. 2008; Digital Curation Centre 2009).

Before such a preservation strategy testing workflow can be implemented, there are five questions that need to be answered:

A. What is the methodology for measuring information loss with respect to syntax, semantics and pragmatics of information?
B. What set of objects constitute a suitable dataset for testing preservation strategies?
C. What are the potential strategies for preserving the information?
D. What risks of information loss are we trying to prevent with these strategies?
E. How do we define acceptable information loss?

These are all challenging questions. The study of significant properties in Sections 3.3 and 3.4 was conducted with the intention of providing us with a partial answer to A. For example, one way of measuring the success of preservation strategies for weblogs would be to show that the categorised contents associated to each data model component are intact, and, the content characterised (using a tool such as FITS[106]) before and after the preservation action results in identical or similar values for the properties listed in Sections 3.3 and 3.4.

This, however, could be viewed more as a methodology to test the preservation of syntax.It may be insufficient to show how the information would change on a macro level at the user interface. The test does not inform us how each action would affect the semantics of the information and it does not inform us how the information would be rebndered in different software/hardware environments. It is not clear at all whether it would meet the pragmatic needs of a designated community. There are also problems of scalability involved in such processes (see Section 4.2.3).

As an approach to obtain an answer to B and a first step to developing a workflow for A, we suggest that we need to understand weblog complexity as defined by the number of potential dependencies the blog has on other resources and blogs. These dependencies can be characterised by the software/hardware environment on which the pages are reliant for instatiation and creation

[106] http://code.google.com/p/fits/

(browsers, operating systems, web servers and blogging platforms), the number and type of content embedded within the pages (e.g. text, image, audio, video), the number of interactive features that are provided within the page (e.g. forms and widgets), the number of references the webpage makes to other webpages (i.e. hyperlinks) and number and depth of topics that are discussed and shared with other blogs. These are all directly related to risks of information loss (see Section 4.2), a view the question C by proxy.

Characterising these features may seem, at first, like a daunting task. However, a rough idea of all these features can be extracted easily from the webpage itself. For example, HTML doctype declarations, range of tags and attribute fields can provide some information on software/hardware dependencies. And traces of blogging platforms can often be found within HTML tag attribute field values (e.g. the term "wp-content" is often indicative of a WordPress blog). The formats and object types used in the blog are also represented through the distribution of file format extension patterns and use context in the form of the associated HTML tags

None of these give a precise picture of the range of dependencies placed on a webpage, but, intuitively speaking, showing evidence that the repository is robust and scalable with respect to increasing variety of these features seems like the way forward to providing evidence that the preservation strategy is effective.

Many of these features are directly related to the notion of a blogging community (e.g. network structure as represented by density of hyperlinks). This implies that, showing preservation strategies to be effective with respect to weblogs that are characterised by similar features leads to systems that better serve selected communities. The extracted features also serve as contextual information, and technical provenance information associated to the blog. In summary, by studying usage of various elements in the weblog while it is still active, we can establish a explicit workflow for:

- **Selecting representative datasets for preservation strategy testing:** The ability to measure and compare complexity is crucial for preservation strategy testing results to be transferable across different organisations.
- **Extracting use context metadata for selected objects:** it highlights the use of different types of objects and formats in the webpage contexts (e.g. images used for mathematical formulas in-line).
- **Estimating the level of complexity that needs supporting in the system (risk assessment):** it helps us to estimate the level of complexity (e.g. scope of object types, format types, and structural constraints) involved in carrying out preservation processes and to determine whether it is scalable and how to make it scalable.

The most challenging question, however, in the above list, is D. It is still an open question to be answered. We have made some informed recommendation in this report but the multi-faceted nature of weblogs makes it difficult to arrive at a definite conclusions. There will more discussion on this topic at the end of this chapter.

In the next section (Section 4.1) we review previously suggested preservation strategies. This will be followed by a discussion of risks in various types of information loss (Section 4.2). In Section 4.3, we present our analysis of four dataset consisting of blog and non-blog pages for features of complexities. We will then end the chapter with some conclusions in Section 4.4.

4.1 Revisiting Preservation Strategies

In this section we list the mainstream approaches to digital preservation that have been suggested within the last couple of decades. This list is not meant to be exhaustive: it is meant to capture the general trends that have generated active discussion. It should also be mentioned that, the strategies

discussed are not intended to be exclusive of each other. In fact in most digital repositories, it is expected that several of these strategies will be adopted in parallel.

Encapsulation and metadata attachment

The objective in this case is to attach the information necessary to interpret and access the bitstream object with the object itself. In its extended form the information could include an executable program that can interpret the bitstream.

Technology museum

This approach involves preserving the versions of software and hardware deemed necessary to access the target information. Some have (for example, the PRADIGM project[107]) commented on the impracticality of this approach (on the basis of storage space for hardware, legal implications with respect to older licenses, feasibility of maintaining older technology that degrade and for which support diminishes, and lack of documentation).

Digital archaeology

This is a methodology for recovering data from objects created using formats, software and hardware that are obsolete. This approach has been largely disregarded as "not a pro-active and preventative approach"[108]. However, if methods of data extraction from unknown formats are studied before the technology becomes unavailable, understanding the technology on a forensics level of accuracy could help to establish an approach for recovering data when the target technology is no longer supported.

Migration

This involves copying or converting one digital object format to another format. Usually, the aim behind the approach is to convert a digital object to a more accessible format before its native format becomes obsolete.

Emulation

The aim here is to use current technology to simulate the environment in which the object was executed. In this approach, there is no change to the object itself. The focus is shifted to the hardware and software environment of the object.

Retargeted binary code & binary translation

This method is related to binary translation, where code written for one system is re-written for another, or code written in one programming language is translated into another language.

Replication

The aim in replication is to make many copies with the vision that it is less likely that all copies will suffer loss at the same time. Projects such as LOCKSS work on this principle by coordinating many copies to repair damaged copies.

Refreshment

[107] http://www.paradigm.ac.uk/workbook/preservation-strategies/selecting-other.html
[108] Ibid.

Whereas *migration* places emphasis on the inaccessibility to objects due to obsolescence, refreshing puts more weight on degradation of information caused by media deterioration and bit rot. Most repositories perform some form of refreshment.

Printing to paper

This approach has now all but disappeared. The objective here was to print out the digital information on paper with the recognition that paper is more likely to survive as it does not depend on rapidly changing technology.

Normalisation

This approach aims to support preservation by converting the objects in the repository to a few select formats for which the repository is able to provide continued support. It is a type of pre-emptive migration methodology.

Standardisation

This can be thought of as a special case of normalisation where the formats within the scope of the repository are aligned to conform to a recognised standard, e.g. ensuring all web pages held within a repository are validated to conform to the XHTML 1.0 standard.

Fixity check

While this process is not usually named as a preservation strategy on its own, it is an integral part of almost any information system, not only for preservation of information, but as a security measure to affirm that the information received did not suffer any errors or contamination during transmission.

In selecting preservation processes, preservation strategies are usually expected to specify: a) formats that will be supported by the repository, and how will they be supported, b) normalisation processes that will be carried out c) standardisation processes that will be carried out, d) migration processes scheduled, e) emulation support that will be provided, f) support software, hardware, associated manuals and data extraction tools that the repository will be collecting and maintaining, g) encapsulation and metadata attachment provided as part of the repository information package, h) selection, verification, and validation methods used for quality assurance. We will return to these questions regarding a) –f) at the end of this chapter. Questions relating to g) will be addressed in Chapter 5 and those relating to h) will be discussed as part of general repository workflow presented in Chapter 6. The final recommended strategy will be summarised and refined in Chapter 7.

4.2 Risk of Information Loss

4.2.1 Missing Links and Incorrect Substitution

The loss of links on the Internet has been studied by several people in the area of web analytics (e.g. Berners-Lee 1998; Bar-Yossef 2004; Gomes 2006). The problem is not always that the information has been deleted but that the URI for the resource has been changed but corresponding links may not have been updated (cf. Gomes 2006). Also, in some cases, links are distributed incorrectly and need to be fixed (see blog post by Jesper Rønn-Jensen[109]). Loss of information that result this way is mostly due to human management problems: keeping track of links and understanding their decay is a big concern. It is partly in response to this type of problem that the Internet Archive started

[109] http://justaddwater.dk/2007/10/05/blog-software-eliminates-user-errors-and-linkrot/

harvesting snapshots of the web. More, recently, there have been other efforts, most notably, perhaps, by the Memento project who have been filling the gap of missing resources by locating an instance of the resource closest in time period and replacing the resource with its *best copy* (with respect to when it was produced).

Data analysis carried out on the four datasets described in Section 4.4 alone showed that approximately 4-33% (depending on the dataset) of the pages could not be retrieved from the originally published URIs. The smallest dataset consisting of URIs recently recommended for categorised content showed the lowest decay, while the general weblog URIs collected in 2009 showed the greatest decay.

Although, we have discussed information loss with respect to missing links to resources above, there are other types of information loss resulting from the incorrect information being provided: for example, the replacement of information with information from an approximately similar time period could lead to conveying incorrect information, especially if the information was intentionally deleted for an explicit purpose. In fact, this type of confusion arises frequently in data management: different versions of scripts, software and datasets are reported along with scientific results on numerous occasions. The expectation would be that similar phenomena would ensue in the web context.

4.2.2 Premature Decisions in Selection

A lot of the current studies regarding profiling formats and technical and semantic aspect of material from the web are conducted on records that have been *uniformised* and packaged as an archival information package, and stored in the institution (for example, the recent study by the British Library[110]). As mentioned at the beginning of this chapter, there is also the question of tracing the digital finger print of information creators: just as handwriting may not be vital for the semantic interpretation of textual entailment, but may carry crucial evidence in the context of authenticity, accountability, tracing history, and other non-textual semiotics, the variations that exist in how we disseminate information (the technical details) may convey information that will allow us to evaluate the authenticity, integrity and usability of the information. It seems, therefore, crucial to examine variations in technical detail (beyond format profiles) to map the correlation between the technical characteristics of the object and the community that produced the object.

The key to dealing with the diversity and changing URIs of web pages is to turn the diversity to our advantage. That is, the mistakes that people make in the construction of their web pages, the preferences they have for selected media types and formats, the tags they choose for describing the content in attribute fields, and the location within a network that they position themselves serve as evidence for characterising the webpage, determining its integrity, completeness, reliability and authenticity.

4.2.3 Inability to Provide Sufficient Preservation Support

Problems of Scalability

The issue of scalability is currently attracting a lot of interest in digital preservation. In developing approaches to preserving information produced online, it is becoming clear that the large volume of information that has to be processed for metadata extraction, validation, identification, and characterisation raises serious concern about whether these processes can be carried out in time, and, consequently about risks in the loss of information that this may entail. If the preservation processes are too complex to be performed to meet the demands of growing web information then the risk to adequate preservation support would be considerable.

[110] http://britishlibrary.typepad.co.uk/webarchive/2012/08/analysing-file-formats-in-web-archives.html

The international internet preservation consortium (IIPC) conducted a test in 2009 (Long 2009) using UK National Archive's DROID[111] and their own format identifier Lister to test a range of preservation processes in migration, emulation, format identification and digital object characterisation. They tested these on a large slice (over 18 million files) of the collection at the Australian National Library PANDORA web archive[112] and found that Lister crashed in five hours and DROID could not complete the task because of a power outage (they estimated that the task would have completed in 42 days). With respect to emulators, they found that, in some cases (e.g. in the case of the BOCHS emulator), the emulator could only handle a smaller sample that could be inserted in the disk image file. Another issue they found with emulators is that often audio and video files do not reproduce properly. They also tried to perform a migration of all Joint Photographic Experts Group (JPEG) images to Portable Network Graphics (PNG) images to be followed by updated links to reflect the change. They found this to be an impossible task on the large dataset, estimating that, just to find all the images, would take over 40 days. They, however, on a smaller dataset, succeeded in identifying and migrating 25,000 JPEG and GIF images to PNG, subsequently updating the links to reflect the change. The migration took 52 minutes followed by 3 days to update the links. The time required for identification of images was not reported.

The scalability issue was also apparent in the small BlogForever experiment conducted using the File Format Tool Set[113] (FITS) which wraps several tools together (DROID[114], JHOVE[115], EXiftool[116], National Library of New Zealand Metadata Extractor[117], FFident[118], and File Utility[119]). On a initial test set of 8 files, the characterisation process (format identification and metadata extraction) took 1.5 minutes and, on a larger set of 500 JPEG images, 500 Portable Document Format (PDF) files, and 505 Hyper Text Markup Language (HTML) files, the process took 4, 11.5, and 32 minutes, respectively. Given that these sizes are microscopic compared to the expected sizes of a web archives, this time constraint is prohibitive. Admittedly, this experiment was carried out on one personal computer running Windows 7, with an Intel Core i5, 2.5 GHz processor and 4MB RAM. However, it is not usually expected that running a characterisation process on 8 files should require massive computational power.

Another challenge with respect to scalability is the trade-off that exists between accuracy and the time it takes to identify and characterise the file as well as the amount of metadata that is expected to add extra storage to the web pages. Currently, an average weblog home page contains around 30 images (see further discussion weblog home pages in Section 4.3). A small collection of 100,000 web pages can be expected to contain 3000,000 images. In our experiments with the FITS, we found an increase of 3.2 MB per 500 JPEG images assigned to metadata. This would imply a 19.2 gigabyte of metadata for the image alone. This estimate discounts additional metadata required for the HTML itself and other embedded media.

The SCAPE project has been advancing research in the area of scalability with respect to preservation process environments in the last year (see their report on characterisation[120]), with a move to integrate the Hadoop[121] architecture for distributed computing with Apache Tika[122] and DROID. While the project has also developed a workflow for preservation processes in the context

[111] http://droid.sourceforge.net/
[112] http://pandora.nla.gov.au/
[113] http://code.google.com/p/fits/
[114] http://droid.sourceforge.net/
[115] http://hul.harvard.edu/jhove/
[116] http://www.sno.phy.queensu.ca/~phil/exiftool/
[117] http://meta-extractor.sourceforge.net/
[118] http://web.archive.org/web/20061106114156/http://schmidt.devlib.org/ffident/index.html
[119] http://unixhelp.ed.ac.uk/CGI/man-cgi?file
[120] http://www.scape-project.eu/wp-content/uploads/2012/05/SCAPE_D9.1_SB_v1.0.pdf
[121] http://hadoop.apache.org/
[122] http://tika.apache.org/

of web content, this does not yet address the scalability issue, and the project did not examine more complex tools that combine several approaches and/or processes such as FITS and JHOVE2[123]. Characterisation can only happen after all instances different embedded objects have been identified from the harvested content, and also must be followed by link updates (see IIPC report[124]). The reason FITS was chosen for consideration was its integrated ability to output the characterisation in widely adopted metadata schema formats.

Problems of Complexity

In addition to scalability issues, web information comes with challenges of complexity in data management. The purpose here is to highlight the issues information loss that may arise, as a result of browser and data dependencies, in data management, say, for example, should the repository decide to perform regularisation (e.g. opting to accept and keep material in the repository using one standardised format), generate selected formats for access (e.g. limiting formats for end-user access copies), carry out migration of material (e.g. to prepare for format obsolescence), employ emulation of systems.

In particular, one observation that may *disappoint* the web archiving community is that web browsers, formats and the way pages are authored in the web environment are diversifying rather than becoming standardised. For example, every time a new versatile mark-up language is developed (say, for instance, HTML5), the previous mark-up invades the authoring of the page; just as long as it renders for the author (regardless of whether or not it passes validation), it makes its way into the Internet. This is a result of several different effects: for example, authors might use tools that automatically create HTML, some parts of which, the author then modifies using newer mark-up standards. Scenarios of this sort are endless.

Browsers

The figures in Table 3-13 show the market share of the various browsers between 2008 and 2011 (shown as a percentage of the total). This data is from Royal Pingdom[125], a blog that has been posting notable statistics (e.g. "internet 2008 in numbers"[126]) concerning information on the Internet for some years now.

Table 4.2-1. Browser market share change over four years (data from royal.pingdom.com).

Browser	2008 (%)	2009 (%)	2010 (%)	2011 (%)
Internet Explorer	69.8	62.7	46.9	39
Firefox	20.7	24.6	30.8	25
Safari	7.2	4.6	4.8	6
Chrome	0.9	4.5	14.9	28
Opera	0.7	2.4	N/A	N/A
Netscape	0.5	N/A	N/A	N/A
Other	0.2	1.2	2.1	2.0

The figure shows the majority share of Internet Explorer declining and the market share of newcomers like Chrome steadily increasing to show a non-majority spread in 2011. In fact, according to a report on Wikipedia[127] the market share of Internet Explorer and Firefox are now at a

[123] https://bitbucket.org/jhove2/main/wiki/Home
[124] http://droid.sourceforge.net/
[125] http://royal.pingdom.com
[126] http://royal.pingdom.com/2009/01/22/internet-2008-in-numbers/
[127] http://en.wikipedia.org/wiki/Usage_share_of_web_browsers

par. The Wikipedia report, however, also illustrates that a local majority does exist (for example, North America still remains dominated by Internet Explorer, while western Europe tends to use Firefox).

The operating systems compatible with these browsers vary as well. For example, Table 3-14 demonstrates that Internet Explorer is only really supported on the Windows operating system, while Google Chrome is supported the three major systems (Linux, Windows, and MacOs) but not BSD or other UNIX variations. And Safari is only supported on Windows and MacOs variations. Opera and Firefox is the only browsers among those listed in Table 3-13 that are still supported on all five major operating systems, Windows, Linux, MacOs, BSD, and UNIX.

On way confirm interoperability between web browsers was suggested through three pages constructed to test: a) conformance with cascading style sheet 1.0 specification (Acid1[128]), b) aspects of HTML mark-up, CSS 2.1 styling, PNG images, and data URIs (Acid2[129]), and, c) Document Object Model (DOM) and JavaScript (Acid3[130]). The results of these tests are dependent not only on web browsers but also the operating system on which they are installed. For example, Internet Explorer installed on Windows may pass the test but not when it is installed on MacOs[131].

The formats supported by each browser vary considerably as well[132] and these also change according to growing online communities (e.g. see recent discussions about support for SVG in browsers[133]).

Web Servers

The figures in Table 4.2-2 present the market share of the various web server platforms between 2008 and 2011 (shown as a percentage of the total). This data, like that for web browsers, comes from Royal Pingdom[134].

Table 4.2-2. Web server market share change over four years (data from royal.pingdom.com).

Server	2008 (%)	2009 (%)	2010 (%)	2011 (%)
Apache	51.0	46.6	59.4	64
IIS	33.8	21.0	22.2	14
Google GFE	5.6	6.0	5.9	9
Nginx	1.8	6.9	6.6	13
Lighttpd	1.6	0.4	0.5	N/A
Other	6.2	19.6	5.4	N/A

The figure shows the majority share of Internet Explorer declining and the market share of newcomers like Chrome steadily increasing to show a non-majority spread in 2011.

Web Page Validation

Web page authors are very poor at conforming to the published standards. Some web authoring mistakes occur because the page was manually edited using a method of trial and error (that is, does the page render in an expected way). On the other hand other mistakes occur because of the authoring tools that we employ in an effort to better conform to standards (for example, it could result from using a tool for a older version of HTML mark-up then copying and pasting the material when the page is updated to a new version or the tool itself not being reliably updated).

[128] http://acid1.acidtests.org/
[129] http://acid2.acidtests.org/
[130] http://acid3.acidtests.org/
[131] http://en.wikipedia.org/wiki/Comparison_of_web_browsers#Acid_Scores
[132] http://en.wikipedia.org/wiki/Comparison_of_web_browsers#Image_format_support
[133] http://royal.pingdom.com/2012/05/15/web-designers-svg/
[134] http://royal.pingdom.com

Pages are often validated by the W3C validation tool[135], modified, and then not re-validated. This also causes discrepancies. It is not even a select few that make each type of mistake. The mistakes are repeated across the Internet in a statistically significant way. This is well demonstrated by studies of tag usage, field usage, and page validity (e.g. Google 2005; Opera 2008). These studies also show that there are very few pages that are valid page (only 4.3% of the dataset Opera 2008 studied was passed WC3 validation). The Google study was used in the development of HTML5, direct evidence that the way people actually author the pages on a technical level influence changes in technology, that is, it might be advisable for digital preservation practitioners to also pay attention to how end-users influence technology, not only how technology and requirements are changing (cf. other technology watch proposals, for example, the watch reports[136] produced by the Digital Preservation Coalition (DPC) and watch workflow for preservation planning produced by the SCAPE project[137]).

The blame for the general failure to produce valid web pages can be often attributed to mismanagement. So, the question is: should we force ourselves to become better managers (very difficult), or create a robust preservation system that can withstand a modicum of mismanagement? Note that only two of the web home pages for the digital preservation projects (listed in Appendix G), and only seven web home pages of web archives (also listed in Appendix G) passed the WC3 validation (at the time of writing this report).

At any rate, it is also important to note that the validation pass of the web page does not guarantee correct rendering of content nor retention of meaning: this is because the syntax and grammar of the page is only loosely coupled with semantics of the page (both for humans and machines alike). And, this is yet another challenge in the preservation of digital information.

4.3 Experiments: Determining Weblog Complexity

In the previous section, we mentioned the obstacles that data complexity and system complexity can present in carrying out preservation processes. However, the complexity of a webpage also results from the variety and inter-related organisation of object types within the web page. Previous work in preservation testing has, so far, done little to address the scalability and information loss risks that result from this level of complexity.

The formatting tags in HTML, while serving as cues for the machine to render it properly on the screen, also serves to tell us what types of objects are included in the page. For example, an <img> tag indicates that an object with the technical format of an image will be included; say, for example, an object with the format of portable network graphics (PNG). Likewise, audio, video, pdf, and java applets can be embedded in the webpage using the <object> tag.

Tag usage analysis, apart from anything else, can shed light on the extent of variation we might expect in a webpage and also help us to estimate the computational complexity involved in characterisation, management, and storage of information related to the web page. The relative frequencies of tags can also help estimate the level of complexity that a webpage structure represents.

The attribute field and value that are associated with the HTML tags can also be extremely informative, even when the fields and attributes are included by error (that is, the inclusion results in an invalid HTML page). A lot of these fields and values are ignored by the browser when they are not relevant: that is, the browser is tolerant to mistakes and displays the expected page anyway.

[135] http://validator.w3.org/
[136] http://www.dpconline.org/advice/technology-watch-reports
[137] http://www.scape-project.eu/wp-content/uploads/2012/01/SCAPE_D12.1_TUW_V1.0.pdf

On the other hand, a lot of these attribute fields and values provide us with extra information that could be useful for the object characterisation process, in some cases rendering the slow and intensive digital object characterisation tools (e.g. FITS) unnecessary.

Another characterising aspect of an *active* blog is the social network to which it contributes. Social network structures have been observed to be correlated to innovative development (Coulon 2005). In this report we present some preliminary observations on webpage link structure (e.g. the density and centrality of blogs in the community network; and how many hyperlinks are self-referential) as an indication of the existence of a blog community and their blogs' potential for supporting community needs. While the link structures of web pages are not comprehensive of the underlying social network structure, we hope to use it as a starting point for encouraging further community involvement in the preservation activity.

Another avenue of investigation we present here is a scoping study of user generated categories and topic tags. This serves two purposes; we are able to determine a) how actively users engage in such activities, b) whether we can leverage their activity to support blog descriptions, and c) whether such activities might be extended to metadata refinement within the repository.

All these aspects of the webpage contribute to a comprehensive view of webpage complexity and leads to a definition of weblog complexity based on technical characteristics, network structure and information sharing behaviour.

The data analysis presented in here shows evidence that the characteristics described above tend to be distinguishing features of blogging communities. In the following sections, we show evidence that there is a striking difference between blogs and non-blogs and between different subcategories of blogs with respect to tag usage, attribute field values, linking behaviour and information sharing. In particular, we argue that testing on large volumes of data is not sufficient to measure the effectiveness of preservation strategies on complex datasets.

4.3.1 Datasets

For the analysis reported here, we have used four datasets consisting of web home pages. The study was limited to home pages, in the first stance, in order to make the study comparable across the datasets without the consideration of different page type variations. The constraint can actually prove to be useful, that is a characterisation of websites based on the home page can easily be replicated without modification across a variety of domains.

The size with respect to each of the examined dataset is presented in Table 4.3-1. The right most column of the table displays the number of distinct HTML doctype declarations used in the dataset. These numbers already suggest that the datasets are very different in character. And further, the low number of distinct declarations with respect to the Mokono dataset is striking when the number is compared to the other datasets, especially in light of the fact that this dataset contains the largest number of pages. In fact, the small Categorised dataset contains more variation than the Mokono and Spinn3r09 datasets, both of which are significantly larger in size.

Table 4.3-1Datasets used in preservation strategy testing experiment.

Dataset Name	Number of Distinct URLs	Number of Distinct Doctype Declarations
Categorised	31690	122
Mokono	312,208	20
Spinn3r09	223,145	80
ClueWeb09	214,952	1420

All the pages for this study were re-harvested during July of 2012, to minimise differences that might arise from pages that are harvested over a long time period.

The first dataset (Categorised) consists of a small set (31690 distinct URLs) subcategorised into sixteen subject and organisational categories. The pages were either collected from blogrolls of the home pages of individuals in the corresponding area or through blog searching portals such as Technorati[138], mathblogging.org[139], scienceseeker.org[140], scienceblogging.org[141] and independent fashion bloggers[142] (ifb). The individuals in selected areas were found by selected those with the top reputations on question and answering forums: StackOverflow[143] for computing science, and MathOverflow[144] for mathematics.

The second dataset (Mokono) of blog home pages were acquired using 344,953 URLs that were provided by mokono-populis[145]. The home pages were crawled using the urlib2 library of python: out of the 344,953 URLs, requests for 312,203 URLs returned a status code of 200 and were used in the analysis.

The third dataset (Spinn3r09) consisted of a random sample from the Spinn3r.com weblog dataset released at the International AAAI Conference[146] on Weblogs and Social Media (ICWSM) in 2009. While the URLs are from 2009, the home pages were harvested again, in July 2012, for the purpose of the study reported here. The dataset consists of 223,145 blog home pages.

The fourth dataset (ClueWeb09) is from the ClueWeb09 dataset[147] created to support research in information retrieval. The original dataset consists of 1 billion web pages. The study in this report is based on a sample of 214,952 home pages. This dataset is not limited to weblogs nor does it exclude weblogs. Some of the pages are expected to overlap with the other datasets. The overlap was allowed to simulate the distribution of web pages that might be found within a general repository of web pages.

The subcategories of the Categorised dataset, number of webpages and their source is listed in Table 4.3-2. Not all of these are blog pages: only where the term “Blog” has been explicitly included should the pages be considered blog home pages. Some of the 3169 URLs in the Categorised dataset are repeated across several of the sixteen subcategories. The overlap is a result of the same pages being recommended within different subject areas. The repetitions were not removed for the study because part of the analysis is to determine similar and different aspects across different blogging communities.

Table 4.3-2 Subcategories of web pages included in the data analysis (internal reference labels, to be used interchangeably with the category name in this report, have been assigned to each subcategory - indicated in the parentheses).

Subcategories	Size	Source
Architecture Company (Arch0)	27	National Building Specification
Computer Science Blog (Comsci10)	41	StackOverflow
Information Technology Blog (Comsci30)	138	Technorati IT Category Search

138 http://technorati.com/
139 http://www.mathblogging.org/
140 http://scienceseeker.org/
141 http://scienceblogging.org/
142 http://heartifb.com/
143 http://stackoverflow.com/
144 http://mathoverflow.net/
145 http://www.populis.com/de/
146 http://blog.spinn3r.com/2008/10/spinn3r-sponsor.html
147 http://www.lemurproject.org/clueweb09.php/

Subcategories	Size	Source
Entertainment Blog (Entertainment10)	110	Technorati Entertainment Category Search
Fashion Blog (Fashion0)	164	Independent Fashion Bloggers
Fashion Company (FashionInd0)	61	http://www.smashingmagazine.com/2009/03/12/showcase-of-beautiful-fashion-websites/ and comments
Funding Councils (Funding0)	51	Search based on personal knowledge
Game Blog (Game0)	7	A PhD student in Games
Government (Government0)	572	http://www.politicsresources.net/official.htm
Health Blog (Health0)	130	Technorati Category Search
Mathematics Blog I (Math40)	110	Fields Medalist Terry Tao's Blog http://terrytao.wordpress.com/
Mathematics Blog II (Math60)	552	Mathblogging.org
Music (Music10)	70	Technorati Category Search
Politics (Politics0)	107	Technorati Category Search
Science (Science0)	1071	Scienceseeker.org and scenceblogging.org
University (University0)	100	http://www.guardian.co.uk/news/datablog/2012/mar/15/top-100-universities-times-higher-education

4.3.2 Variation of HTML Versions

The *Doctype* (definition of how this declaration is used can be found at the W3Schools tutorial[148]) declaration which specify the HTML versions and flavours was initially examined across the dataset. The majority of pages (that is, more than half of each dataset) were declared as a version of XHTML. However, a large number of pages do not bother with a declaration at all (see Spinn3r09 result in Table 4.3-3). It seems that web authors tend not to declare their document types explicitly.

Table 4.3-3 Web page document type percentage within each of the four datasets (percentages are approximate).

Doctype	Categorised (%)	Mokono (%)	Spinn3r09 (%)	ClueWeb09 (%)
No declaration at all	3.9	0.00032	44.44	14.07
"doctype html" or "doctype html public"	31.42	0.00704	1.77	12.47
XHTML	59.09	99.98	51.82	55.19
MOBILE	0.057	0	0	0.072
OTHER HTML	6.953	0.00672	1.97	19.82

Table 4.3-4 Percentage of web pages declaring selected HTML flavours (percentage is in relation to entire dataset so does not add up to 100)

HTML Flavour	Categorised (%)	Mokono (%)	Spinn3r09	ClueWeb09
Transitional	43.34	99.98	47.62	51.67
Strict	18.26	0.003	5.59	17.1
Frameset	0.31	0	0	0.47

In Table 4.3-4, we have displayed the percentages of pages declared to be "Transitional", "Strict", and "Frameset". It is clear that most pages have opted for the transitional flavour which implies that

[148] http://www.w3schools.com/tags/tag_doctype.asp

there will be lots of deprecated tags and attribute fields appearing in these pages. We will see in the later sections that this is indeed the case.

In the case of the Categorised, Spinn3r09, and Mokono datasets, the pages which have been declared XHTML are also declared to be version 1.0 with only an exception of six instances in the categorised dataset which are declared as version 1.1. Most of the pages that have claimed to be XHTML in the ClueWeb09 dataset also claim to be version 1.0. However, the ClueWeb09 dataset declarations are messy: they also mention 84 instances of version 1.1, six instances of version 4.01, four instances version 4.0, two instances of version 11.0, and one instance each of versions 1.00, 1, .0, and 2.0. Some of these, of course are invalid version numbers for XHTML: checking the source pages show that they, nevertheless, have been declared as XHTML in this way.

In the case of those pages declared HTML, the biggest share declare their version to be 4.01 and 4.0 (more 4.01 than 4.0) with a few additional instances of version 3.2. The ClueWeb09 dataset again shows a confused outcome with some remaining pages declaring versions 5.0, 5.01, 4.1, 1.0, and 2.0.

One immediate observation can be made on the regularity of the pages in the Mokono dataset: the pages in the dataset have very few variations in their declarations. This makes it perhaps an easier dataset to manage and preserve but suggests that it may not be suitable as a dataset representative of the blogosphere, and as a dataset for use case study (planned to be carried out in BlogForever WP5) or preservation strategy testing. The regularity of the pages in the Mokono dataset will also be clear in the next discussions regarding HTML tag usage.

The declaration of the form "doctype html" is a standard for HTML5[149]: because HTML5, unlike HTML 4.01, is not based on Standard Generalised Markup Language[150] (SGML), DTD location information is no longer required, which is the portion that contained the pointer to the version. We have displayed, in Table 4.3-5, the number of pages with declarations conforming to HTML, XHTML and HTML5 with respect to each of the subcategories in the Categorised dataset.

Table 4.3-5 Number of HTML, XHTML, and HTML5 pages with respect to varying categories.

Domain	HTML	XHTML	HTML5
Architecture Company (Arch0)	2	21	1
Computer Science Blog (Comsci10)	0	33	8
Information Technology Blog (Comsci30)	6	78	54
Entertainment Blog (Entertainment10)	7	74	28
Fashion Blog (Fashion0)	4	100	57
Fashion Company (FashionInd0)	8	31	17
Funding Council (Funding0)	5	41	4
Game Blog (Game0)	1	0	6
Government (Government0)	108	336	28
Health Blog (Health0)	4	73	53
Mathematics Blog I (Math40)	2	81	24
Mathematics Blog II (Math60)	14	313	218
Music Blog (Music10)	12	33	21
Politics Blog (Politics0)	5	73	27
Science Blog (Science0)	16	570	482
University (University0)	18	67	12

[149] See W3Schools explanation at http://www.w3schools.com/html5/tag_doctype.asp
[150] http://en.wikipedia.org/wiki/Standard_Generalized_Markup_Language

The widely varying numbers of pages belonging to each category makes it difficult to make firm conclusions but it is clear that there is move towards HTML5 in authoring weblog pages, especially in the sciences and related subjects (e.g. health and general mathematics). Non-blog pages such as Government, Architecture Company, Funding Council, and University home pages remain predominantly HTML and XHTML based.

4.3.3 Usage of HTML tags across datasets

In this section, we will discuss correlations between different tags, how they relate to the blogging platform and the community that the platform tends to serve. First, in Table 4.3-6, we present the number of distinct HTML tags, and the number of all HTML tags used within each dataset.

Table 4.3-6 Number of HTML tags in each of the four datasets.

	Categorised	Mokono	Spinn3r09	ClueWeb09
Distinct HTML tags	250	185	1011	2774
All HTML tags	2676612	158506533	139176708	145535289

Note that, although the Mokono dataset is the largest of the four datasets, the number of distinct tags used through the dataset is the smallest. This again confirms our observation earlier about the technical regularity of the Mokono dataset. As expected, the ClueWeb09 dataset (consisting of the most variable set of pages) contains the most number of distinct tags.

Given that this is only a small sample of the information available on the web, it is quite striking how many tags are actually in use. While most of these tags are bound to be associated to the layout of the page, others will be associated to images, videos and other embedded content. If we aim to provide the recommended metadata for all of these embedded contents, that alone could pose serious problems in relation to scalability. This is especially daunting when we recall that these pages are only the home pages (no post pages have been included).

In fact, out of the 145.5 million or so tags used within the ClueWeb09 dataset, just over 7.4 million tags are estimated to be <*img*> tags, 44573 estimated to be <*object*> tags, and 112945 estimated to be related to a flash video object. Likewise, just under 5 million of the 139.1 million Spinn3r09 HTML tags have been identified as <*img*> tags, and approximately 2 million of the Mokono dataset HTML tags have been identified as <*img*> tags.

Table 4.3-7 Frequent HTML tags in the dataset (definitions are from www.w3schools.com/tags).

Tag	Definition
<a>	The <*a*> tag defines a hyperlink, which is used to link from one page to another.
** **	The tag inserts a single line break.
** **	The tag defines a division or a section in an HTML document.
****	Renders as emphasized text
<h2>	The <h1> to <h6> tags are used to define HTML headings.
<img>	The <img> tag defines an image in an HTML page.
<input>	The <input> tag is used to select user information.
** **	The tag defines a list item.
<link>	The <link> tag defines the relationship between a document and an external resource.
<meta>	The <meta> tag provides metadata about the HTML document. Metadata will not be displayed on the page, but will be machine parsable.
<option>	The <option> tag defines an option in a select list.
<p>	The <p> tag defines a paragraph.
<script>	The <script> tag is used to define a client-side script, such as a JavaScript.
****	The tag is used to group inline-elements in a document.
****	Defines important text
<td>	The <td> tag defines a standard cell in an HTML table.

<tr>	The <tr> tag defines a row in an HTML table.
** **	The tag defines an unordered (bulleted) list.
<embed>	The <embed> tag defines a container for an external application or interactive content (a plug-in).
<form>	The <form> tag is used to create an HTML form for user input.
<iframe>	The <iframe> tag specifies an inline frame.
<object>	The <object> tag defines an embedded object within an HTML document. Use this element to embed multimedia (like audio, video, Java applets, ActiveX, PDF, and Flash) in your web pages.

Despite the large number of tags, about 22 distinct tags account for 93.79% of the Categorised dataset tag usage, 92.50 of the Mokono dataset tag usage, 87.96% of the Spinn3r09 dataset tag usage, and 90.82% of the ClueWeb09 tag usage. These tags are presented in Table 4.3-7 along with definitions from W3Schools.com[151]. These tags include the top ten most frequently used tags within the four datasets, plus four additional tags (the last four in the list) selected to capture information about media being added into blogs.

Even though the number of distinct tags is largest in the ClueWeb09 dataset, the average number of distinct tags per page is slightly less than that of the other datasets. On average, the number of distinct tags in a page for a ClueWeb09 page would is around 23-27 while that of the other dataset is around 27- 33. The difference may seem small, but, according to a two-tailed t-test with $P<0.05$, the difference is statistically significant. Combined with the fact that the total number of distinct tags in the dataset is largest for the ClueWeb09 dataset, this seems to imply that there are more types of tags being used in blog pages but less individuality in terms of tag usage across pages. This makes sense, since the basic blogging platform is designed to be a pre-formatted template that can be used to start a blog by even those with hardly any technical knowledge.

The relative ratio of the tags (introduced in Table 4.3-7) being used with respect to the four datasets are presented in Figure 4.3-1. The figure shows that, whiles tags such as *<a>*, *<link>*, *<script>* are evenly weighted across the datasets, there are noticeable differences. For example, the ClueWeb09 dataset contains a relatively large number of *<option>* tags, the Mokono dataset contains a relatively large number of ** tags, the Spinn3r09 dataset seems to contain a relatively large number of *<object>* and *<embed>* tags and the Categorised dataset contains a relatively large number of ** and *<iframes>* tags. Notes that there may not be large numbers of instances with respect to these tags and datasets; the observations are relative to the other tags and datasets.

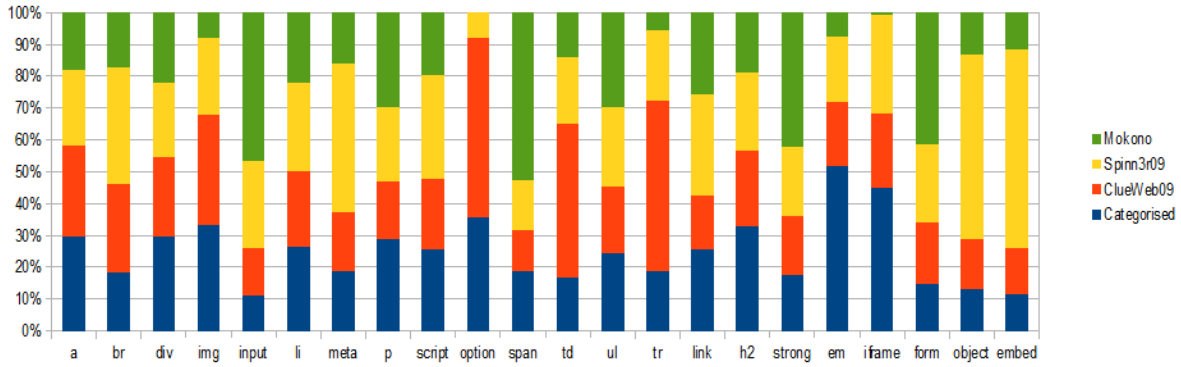


Figure 4.3-1 Relative ratio (y-axis) of selected HTML tags (x-axis) across four datasets (represented in different colours).

As a further example to highlight the differences between the datasets, in Figure 4.3-2, we have presented the number of pages (indicated using a blue bar and subsequent number, followed by the

[151] http://www.w3schools.com/tags/

percentage with respect to sample size in parenthesis) across distinct tag counts (left hand column of numbers) for a random sample of hundred pages from each of the four datasets. The figure shows the pages in the Mokono dataset to be quite similar to each other with respect to the number of distinct tags (that is, not only is 75% of the pages within the range of 24-32 distinct tags, but also more of the pages share exactly the same number of distinct tags), while the distribution across tag counts to be more variable in the other datasets. In particular, 60% of the pages in the ClueWeb09 dataset sample have less than 24 distinct tags.

The tendency of the different datasets to cluster around different frequencies is more clearly visible in Figure 4.3-3. The numbers on the y-axis is the cluster label for each dataset and the x-axis represents the number of distinct tags.

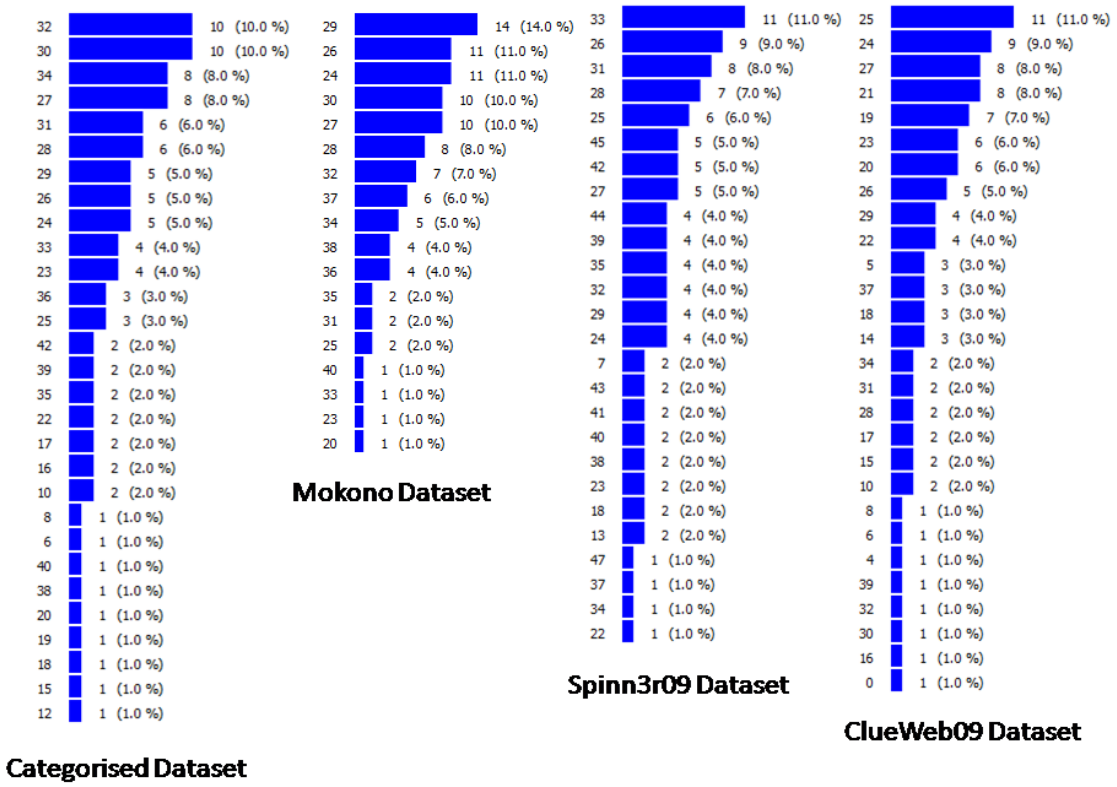


Figure 4.3-2 Number of pages across distinct tag counts in samples of 100 pages from four datasets (left hand column of numbers indicate the count of distinct tags. The blue bar and subsequent numbers indicate the pages with the tag count).

The variation across the sixteen subcategories of the Categorised dataset is displayed in Figure 4.3-4. The keys to the figure are: Fashion Blog (cluster 1 in dark blue), Funding Council (cluster 2 in red), Fashion Company (cluster 3 in light green), Politics Blogs (cluster 4 in orange), Math Blog II (cluster 5 in yellow), Computer Science Blog (cluster 6 in purple), Music Blog (cluster 7 in cyan), Game Blog (cluster 8 in light grey), Entertainment Blog (cluster 9 in light blue), Government (cluster 10 in light brown), Information Technology Blog (cluster 11 in dark grey), Mathematics Blog I (cluster 12 in dark brown), Science Blog (cluster 13 in dark green), Health Blog (cluster 14 in olive), University (cluster 15 in blue green), and Architecture Company (cluster 16 in brown).

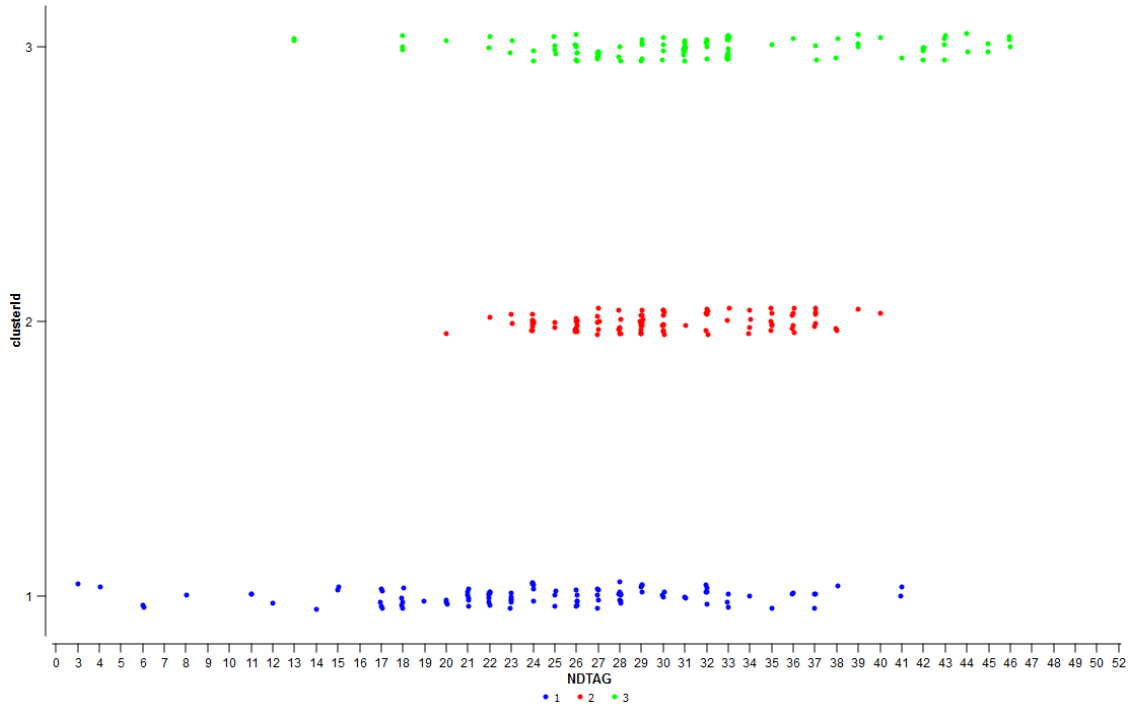


Figure 4.3-3 Number of distinct tags in the datasets ClueWeb09 (blue coloured cluster 1), Mokono (red coloured cluster 2)and Spinn3r (green coloured cluster 3).

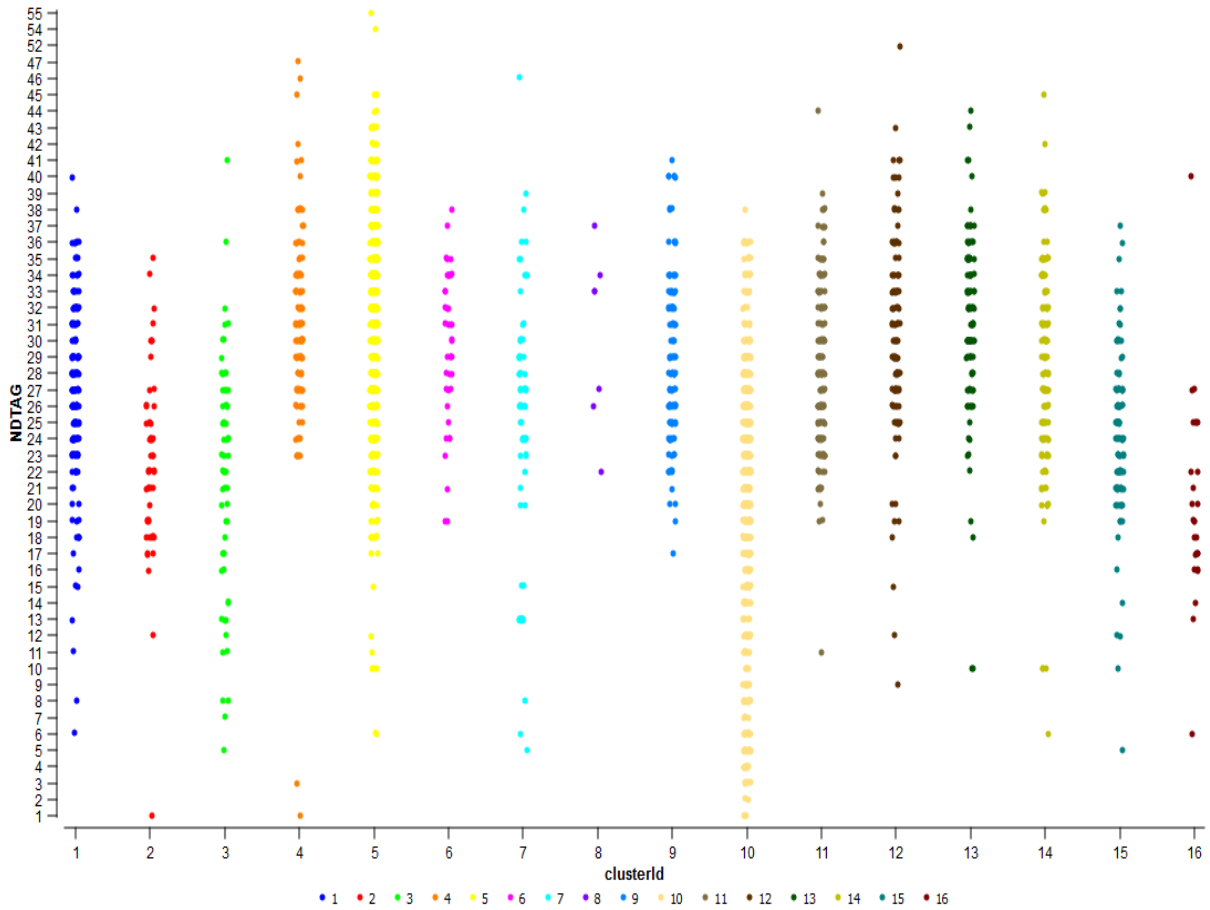


Figure 4.3-4 Number of distinct tags for home pages in 16 domains.

While some of the distributional characteristics displayed in Figure 4.3-4, initially seems to be a result of the data size (for instance, the wide distribution of distinct tag counts for the Government home pages (cluster 10 in light brown), when you compare it to Mathematics Blog II which is of similar size (cluster 5 in yellow) it is clear that it is not just a matter of size. In fact, the pattern suggests that non-blog pages (most distinctively, Fashion Company, Government and University) contain less number of distinct tags on average.

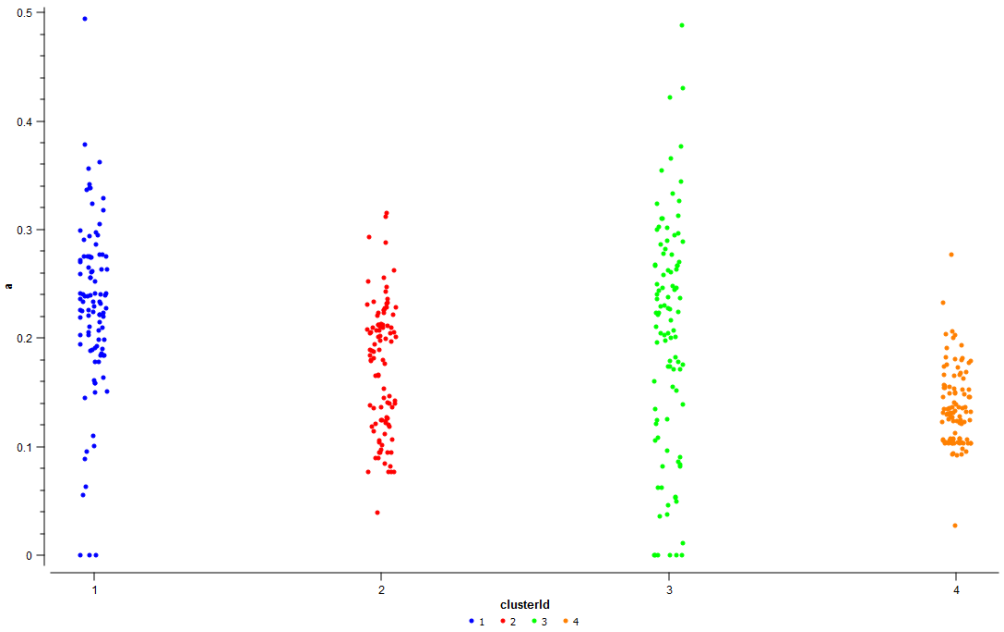


Figure 4.3-5 Relative "a" tag count with respect to all tag counts across datasets Categorised (cluster 1 in blue), Spinn3r09 (cluster 2 in red), Clueweb09 (cluster 3 in green), and Mokono (cluster 4 in orange). The y-axis indicates the ratio, the count of "a" tags divided by the count of all tags.

In addition to the raw numbers of tags, we examined the relative frequency of the 22 tags frequently used in the web page: that is we expressed each page as a vector consisting of the ratio of the count of each tag over the count of all tags used in the web page. The lack of diversity with respect to the Mokono dataset, already observed in Figure 4.3-2and Figure 4.3-3, is also noticeable with respect to the ratio of individual tags. As an example of this tendency, the ratio of <*a*> tag count (the most frequent tag in all the datasets) across samples (of size 100) from the three datasets is presented in Figure 4.3-5. The figure shows that, for most of the URLs in the Mokono sample, 10-20 percent of the tags used are instances of the <*a*> tag. The relative frequencies with respect to “img” tags (in Figure 4.3-6) show similar results. The deviation of the Mokono dataset from all the other datasets is seen consistently across all the 22 tags examined.

We used a hierarchical clustering algorithm on the relative tag frequencies to see whether these clusterings correlate with the choice of blogging platform and/or type of blog. The result with respect 100 random blogs from the Spinn3r09 dataset (Figure 4.3-7) shows that, for example, a lot of WordPress blogs cluster together (the pink area in the figure) and MySpace blogs cluster together (blue area of the figure) on the basis of relative tag frequencies. The large number of MySpace blogs is indicative of the fact that the URLs provided with the Spinn3r dataset was collected in 2009. However, the usage statistics represents how those tags are being used now, as all the home pages were re-harvested during July, 2012.

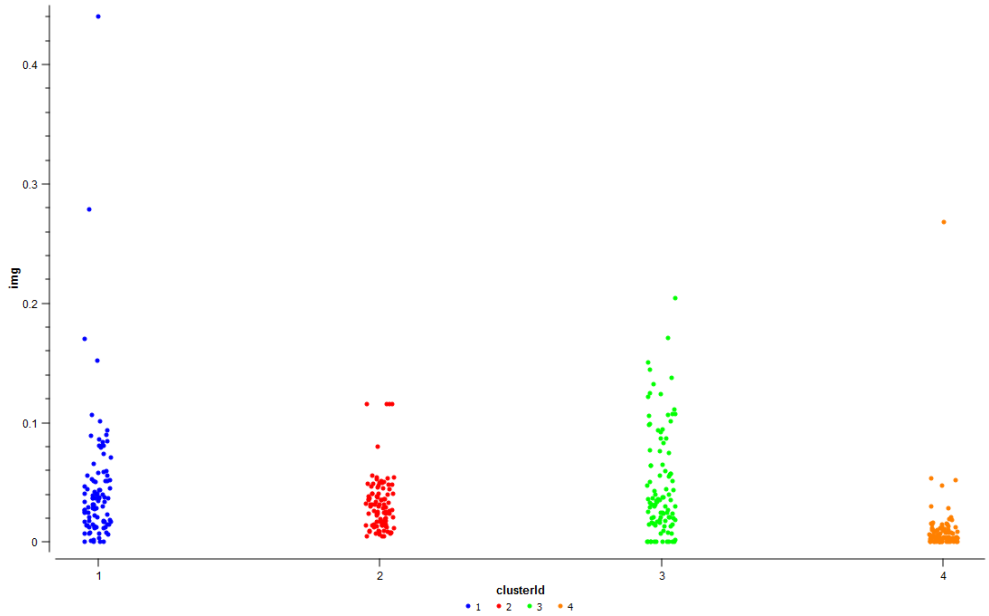


Figure 4.3-6 Relative "img" tag counts across datasets Categorised (cluster 1 in blue), Spinn3r09 (cluster 2 in red), ClueWeb09 (cluster 3 in green), and Mokono (cluster 4 in orange). The y-axis indicates the count of "img" tags over the count of all tags.

The clusters were built through agglomerative hierarchical clustering using the Ward's linkage method[152] for measuring distances between clusters. Ward's linkage method is based on minimising the variance of resulting clusters.

Clusters were also produced on the Politics Blog subcategory (Figure 4.3-8). The interesting here is that there is some subject related clustering visible (e.g. thecable.foreignpolicy.com and shadow.foreignpolicy.com seems related and thinkprogress.org and occupywallst.org to sit close to each other as two left-leaning sites with a similar agenda). Some surprising results may include the fact that "Huffington Post" – which Republicans believe has it in for them – sit so closely with Virginia Right (and the All American Blogger, which is certainly conservative).

The technical level clustering observed with respect to the Spinn3r09 dataset is harder to capture here by only looking at the URLs (this in itself is interesting, in that many of the blogs in the Categorised dataset use their own domain names in the URL rather than using the free default URL provided by the blogging platform provider. Nevertheless some evidence of such clustering is still visible (e.g. the cluster consisting of the two blogs both from "salon.com").

The above results may not show any clear evidence that tag usage is tied to blogging communities, but, the result of applying a self-organising map to the relative tag frequencies of URLs from the sixteen subcategories of the Categorised dataset clearly shows that URLs from the same domains tend to share relative tag frequencies. The visualisation of this result is shown in Figure 4.3-9. The keys for categories displayed in the figure are: Fashion Blog (cluster 1 in dark blue), Funding Council (cluster 2 in red), Fashion Company (cluster 3 in light green), Politics Blogs (cluster 4 in orange), Math Blog II (cluster 5 in yellow), Computer Science Blog (cluster 6 in purple), Music Blog (cluster 7 in cyan), Game Blog (cluster 8 in light grey), Entertainment Blog (cluster 9 in light blue), Government (cluster 10 in light brown), Information Technology Blog (cluster 11 in dark grey), Mathematics Blog I (cluster 12 in dark brown), Science Blog (cluster 13 in dark green), Health Blog (cluster 14 in olive), University (cluster 15 in blue green), and Architecture Company (cluster 16 in brown)

[152] http://www2.statistics.com/resources/glossary/w/wardslnkg.php

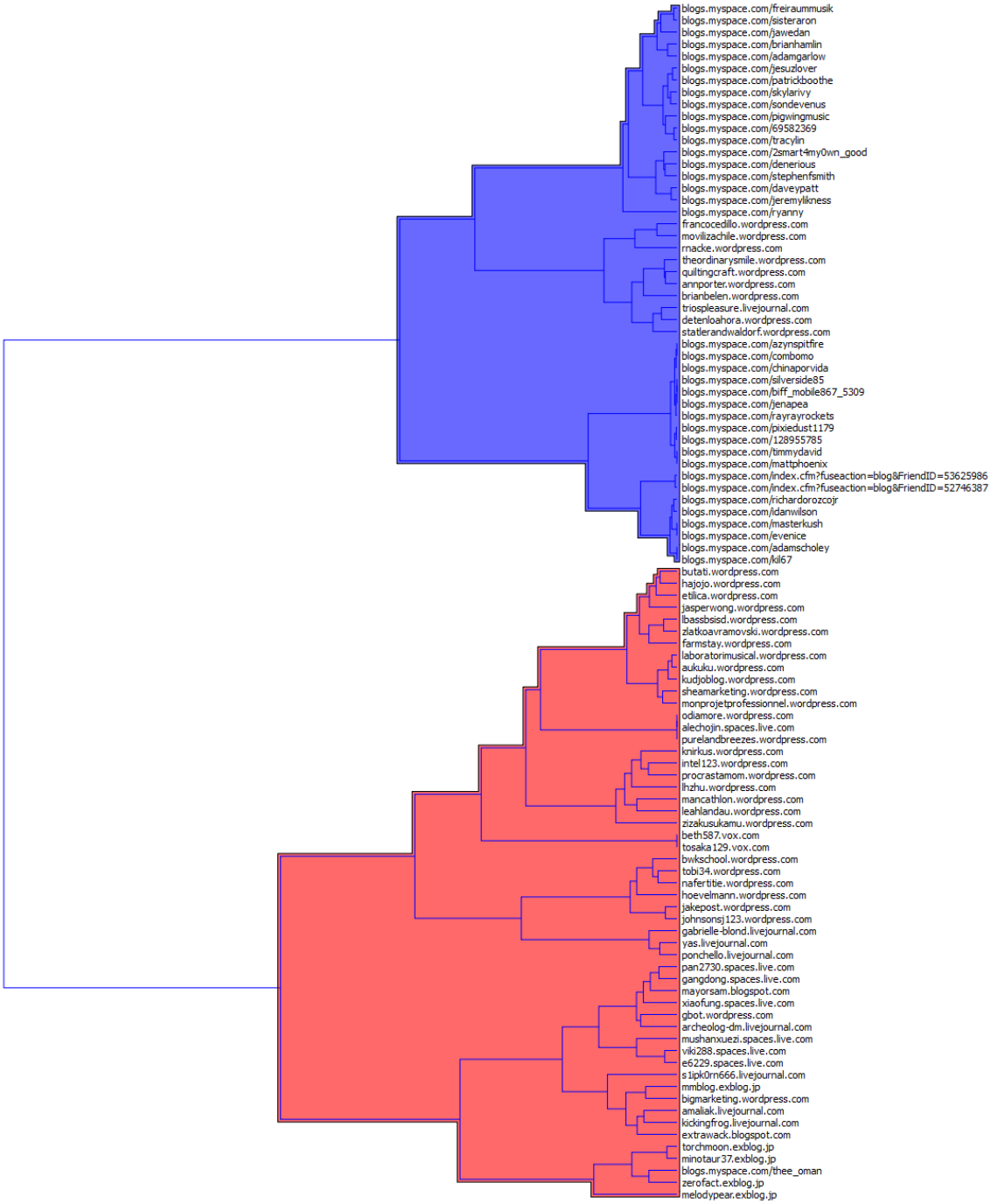


Figure 4.3-7 Hierarchical clustering of URLs according to relative tag frequencies using a random sample from the Spinn3r09 dataset.

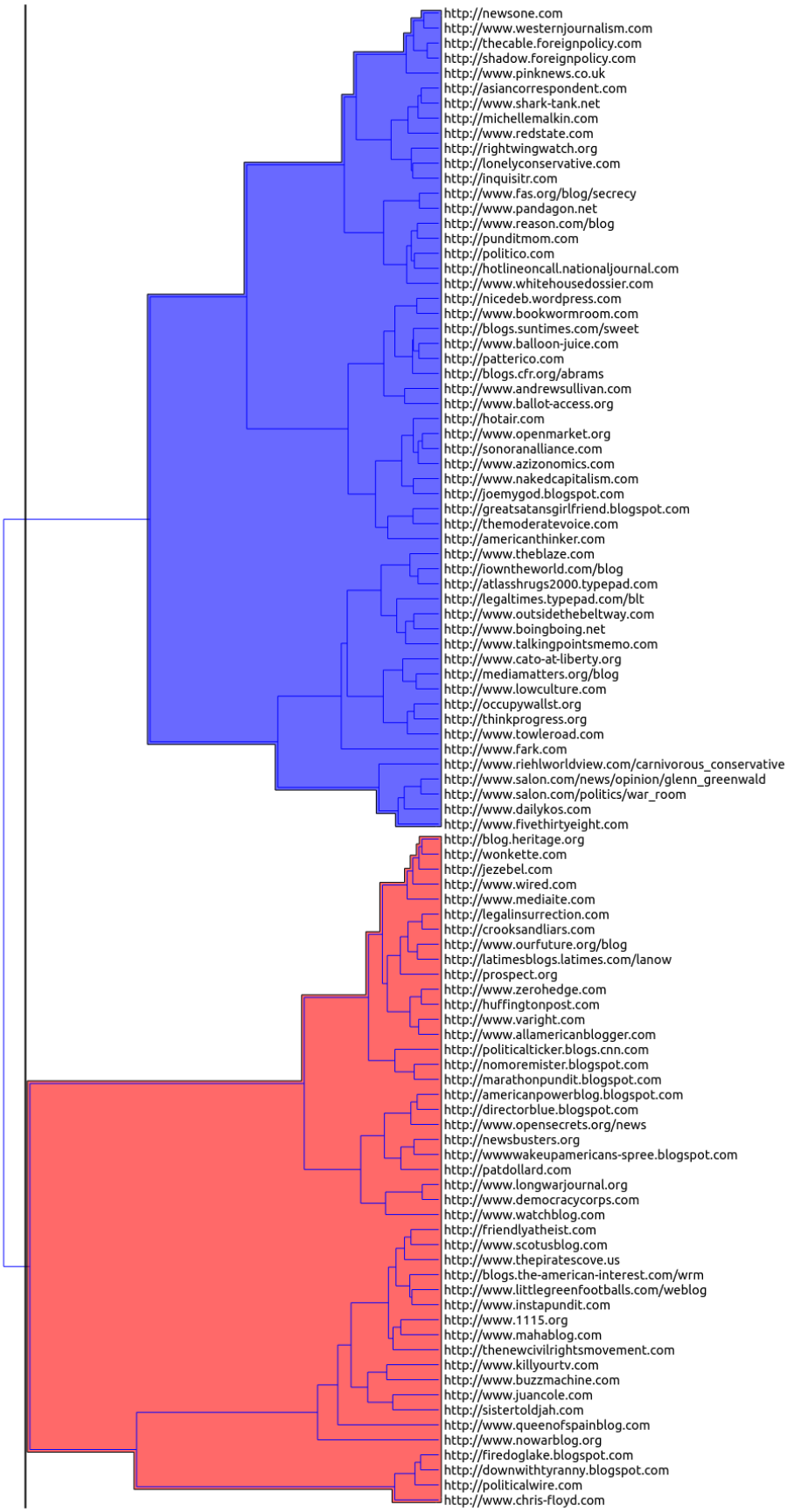


Figure 4.3-8 Hierarchical clusters of URLs from the Politics Blog domain based on relative tag frequency.

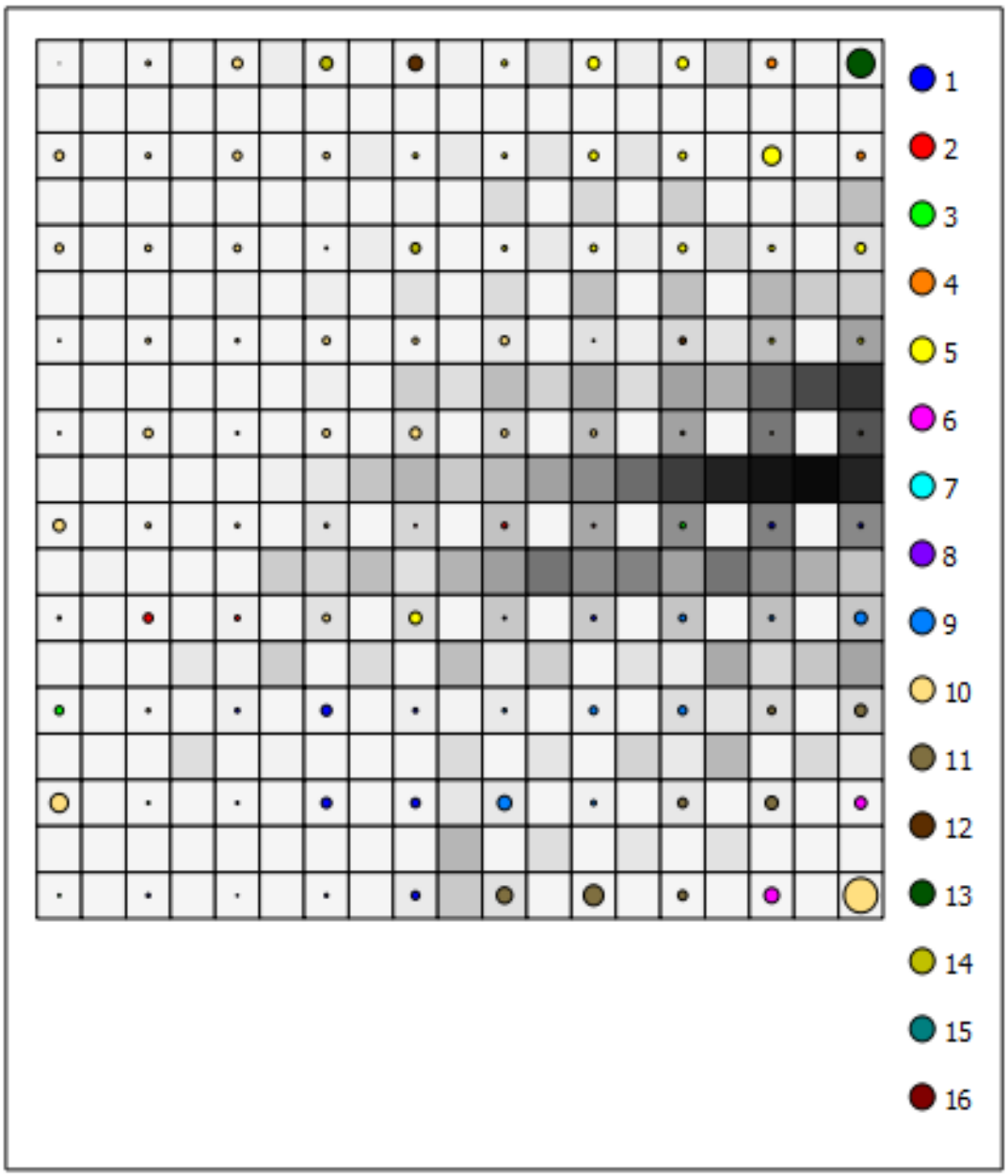


Figure 4.3-9 Self-Organising map showing clusters of subcategories of the Categorised dataset. colours indicate URLs from the same subcategory.

4.3.4 Platforms Adopted by Blogs Across the Datasets

The detection of platforms used in the blog is not straightforward. This information is often intentionally hidden by blog platform providers to discourage hackers. Here we give a rough estimate based on attribute field values that appear within the HTML tags *<link>, <script>, and <meta>* which often contain pointers to resources such as stylesheets, JavaScripts, and background images that are intended to be applied to the entire blog. Tags likely to contain such global parameters tend to be more likely to have a reference to the blogging platforms and/or content management software.

Here we looked for patterns containing the names of 20 well known blogging and/or content management platforms including WordPress[153] (abbrev. wp), Blogger[154], Blogspot[155], Tumblr[156], Typepad[157], Joomla[158], drupal[159], MovableType[160], Pinterest[161], LiveJournal[162], PosterousSpaces[163], Jux[164], Weebly[165], MySpace[166], and SquareSpace[167]. This is intended to give us a rough idea of what platforms might be in use. The result for the Categorised dataset is presented in Figure 4.3-10. A similar analysis of platforms in the Spinn3r09 dataset is presented in Figure 4.3-11.

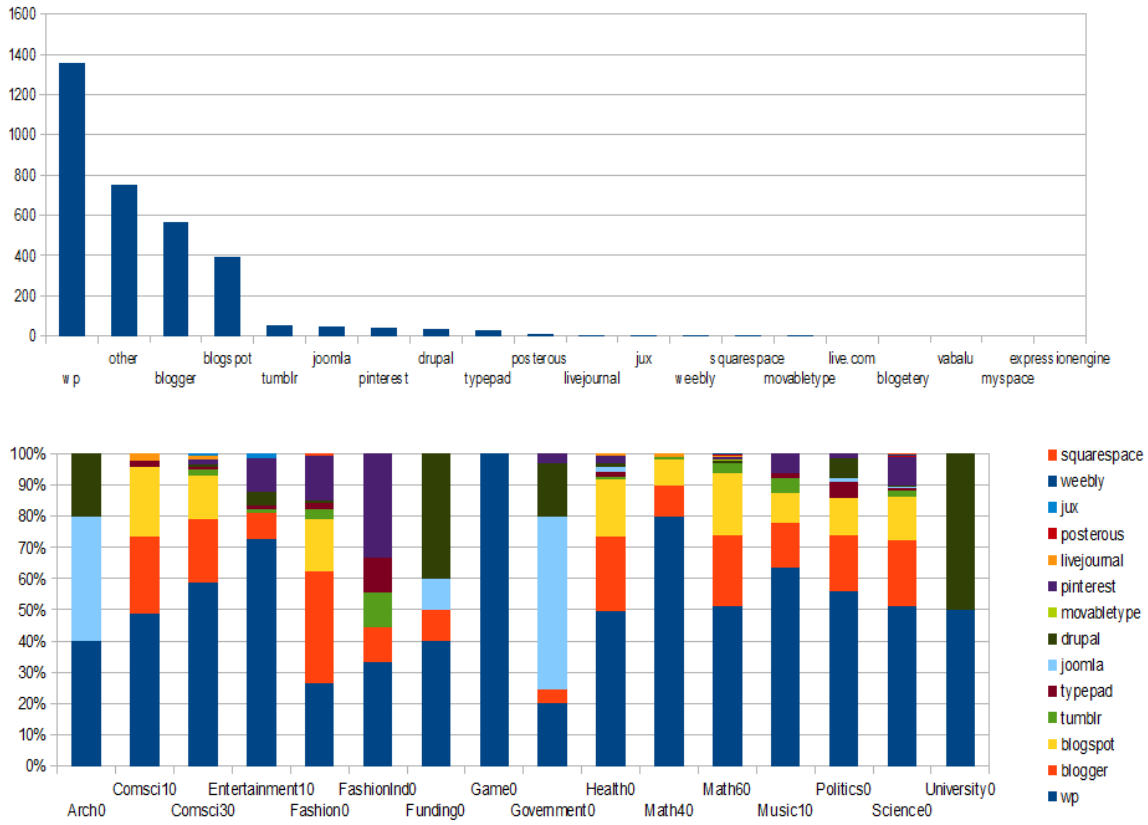


Figure 4.3-10 Estimated numbers of blogging platforms used in the Categorised dataset. Top image shows the distribution in the whole dataset. Bottom figure shows the distribution with respect to each subcategory in the Categorised dataset (each colour in the bottom figure represents a different blogging or content management platform).

153 http://wordpress.com/
154 http://www.blogger.com
155 Accessible through Blogger.
156 http://www.tumblr.com
157 http://www.typepad.com
158 http://www.joomla.org
159 http://drupal.org/
160 http://www.movabletype.org/
161 http://pinterest.com/
162 http://www.livejournal.com
163 https://posterous.com/
164 https://jux.com/
165 http://www.weebly.com/
166 http://www.myspace.com/
167 http://www.squarespace.com/

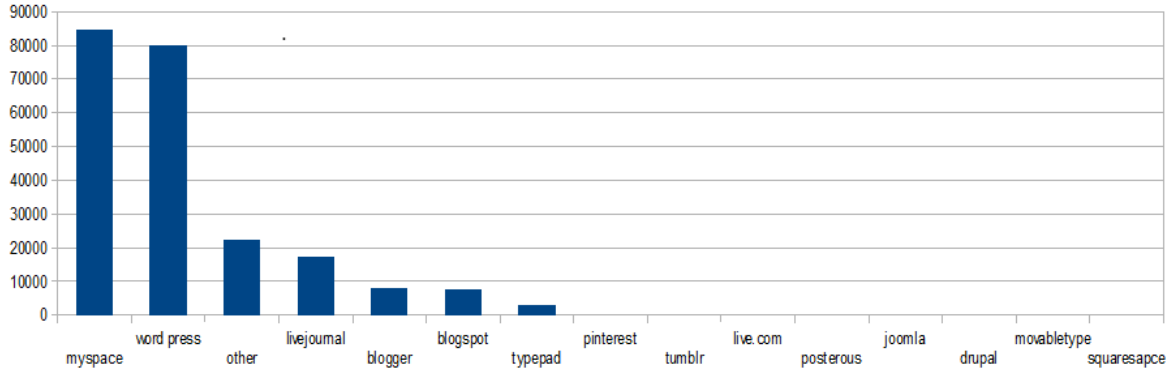


Figure 4.3-11 Estimated numbers of platforms in use within the Spinn3r09 dataset.

There were many pages for which we could not determine an association with any of these blogging platforms. This is especially prominent in the Categorised dataset where nearly third of the pages could not be characterised. Perhaps this is also an effect of the diversification of technologies that we earlier showed (in Section4.2.3) is a currently observable phenomenon. Regardless of the reason, the *non-conformist* tendencies of the Categorised dataset, suggests the technological diversity we can expect.

The distribution for the Mokono dataset was not presented here because this collection consists of blogs that are hosted by mokono-populis[168] and therefore contain very few blogs using these platforms if at all. A quick search shows that there are only 2 references to WordPress within over 80,000 pages from this dataset. This may explain the regularity of the HTML versions and tag usage within this collection that we have already presented in Sections 4.3.2 and 4.3.3.

4.3.5 File Format Extensions Used by Blogs

In the previous section, we examined the correlation between the datasets and tag usage statistics. In this section we present some statistics regarding "file format extensions". The term has been placed in quotation marks because, strictly speaking, what is examined here is not file format extension, but, 5 character patterns in the form ".xxxx" that look like file extensions and that have been found within the attribute field values of the webpage. These patterns have been extracted using the python os.path.splitext module on the HTML tag attribute field values. No specific tag or attribute field was targeted: all attribute values were assumed to have an extension. Consequently, the most frequent pattern returned was the empty string. These were not analyzed here.

There are a lot of patterns that can be extracted in this way. A random sample of hundred URLs from each of the four datasets results in 1179 patterns each from the Mokono, Spinn3r09, and ClueWeb09 datasets. Because there was no post-processing to match terms, many of these pattern are related or are variations of each other (e.g. .jpg and .jpeg; .ico and .ico?).

The objective here is to get an idea of the type of objects and resources embedded and referenced within the page and how these are related to the blogging community. While these do not tell us enough to identify the actual resource or embedded object that the 5 character pattern represents we will show here that the pattern can tell us a lot about the technical aspects of the page and how this relates to the blogging community.

First, we expressed each of the sixteen subcategories of the Categorised dataset as a vector where each of the coordinates are correspond to one of selected 49 patterns (those that are shared across more than 2 URLs in the random 100 URL sample) resembling file extensions (extensions are

[168] http://www.populis.com/de/

displayed in Table 4.3-8). Each coordinate value of the vector was calculated as the number of URLs in the collection that use the corresponding pattern divided by the total number of URLs.

Table 4.3-8 Range of patterns resembling file extensions that have been found to be prolific within the four datasets.

.bmp	.ico	.aspx	.0	.01	.mp3	.stor
.js	.xml	.g	.pdf	.3	.avi	.org
.css	.html	.post	.4	.jpeg	.ogg	.uk
.jpg	.com	.htm	.cfm	.jsp	.nt	.de
.php	.serv	.href	.1	.swf	.cgi	.fr
.png	.valu	.asp	.shtm	.flv	.6666	.be
.gif	.dele	.gete	.2	.mp4	.net	.es

These sixteen vectors representing each subcategory were then clustered using a hierarchical clustering algorithm that uses the Ward's Linkage[169] method to calculate distances between clusters and Pearson Correlation to measure distances between vectors. This clustering method is an agglomerative method, i.e. at first all vectors are assumed to be a cluster of their own, then other vectors are merged into the cluster in a way that the closest vectors are merged into one cluster. The Ward's Linkage method is based on the general idea of minimising the variance within the resulting clusters.

The result of the cluster is presented in Figure 4.3-12. The figure shows a clear division between blogs (clustered in the pink area) and non-blogs (blue area). The blog clusters are also striking. There are clusters that we might expect judging on a semantic and pragmatic level (e.g. Music Blog is clustered with Entertainment Blog; Mathematics Blog clustered with Computer Science Blog). However, the fact that Math40 was grouped with Comsci10, while Math60 is grouped with Health0, Science0 and Comsci30 is also indicative in that the blogs in Math40 and Comsci10 are from individuals with good reputation on StackOverflow[170] and MathOverflow[171], while the blogs in Math60, and the other blogs are the top blogs returned by a the blog search engine at mathblogging.org[172], Technorati[173], and ScienceSeeker[174].

Some of the patterns in Table 4.3-8 are clearly not file format extensions. For example single digit number patterns such as ".0" are likely to the latter part of a version number and patterns such as ".uk" and ".com" are likely to be parts of URLs. There are also four digit numbers often appearing in these patterns. While some of them are truncated font size information (the ".6666" in the table of patterns is such a case), others can be references to preprint server articles. In fact an examination of the Categorised dataset shows that there are 656 citations of articles at arXiv.org[175] in the dataset. The preprint server arXiv.org is used widely in the Mathematics, Computer Science and Physics disciplines, and, now provides trackback functionality for those who want to cite the articles in their blogs. About 37-46% of the URLs in Comsci10 and Math40 datasets have used this service.

The variation of format usage across the disciplines is again visible in Figure 4.3-13. The figures clearly shows that, while there are patterns like ".js" which are shared by most URLs across all subcategories, patterns such as ".gif" widely varies across the categories. Likewise, patterns such as ".mp4" and ".mp3" are only few across all categories and URLs, while ".pdf" is more variable

[169] http://en.wikipedia.org/wiki/Ward%27s_method
[170] http://stackoverflow.com/
[171] http://mathoverflow.net
[172] http://www.mathblogging.org
[173] http://technorati.com/
[174] http://scienceseeker.org/
[175] http://arxiv.org/

across domains. The figure also shows that patterns such as ".valu", ".serv", ".dele" are patterns specific to blogs as very few non-blog category pages seem to contain them.

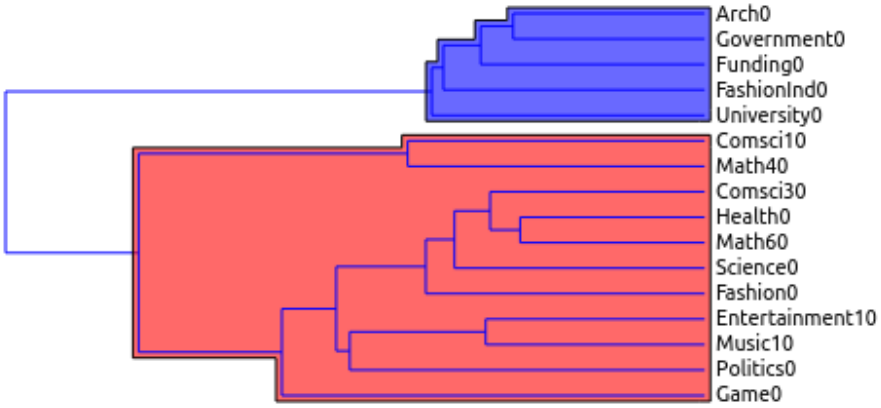


Figure 4.3-12 Hierarchical cluster of subcategories from the Categorised dataset. Note that the domains in the blue area are represented by non-blog pages and those in the pink area are represented by blog pages.

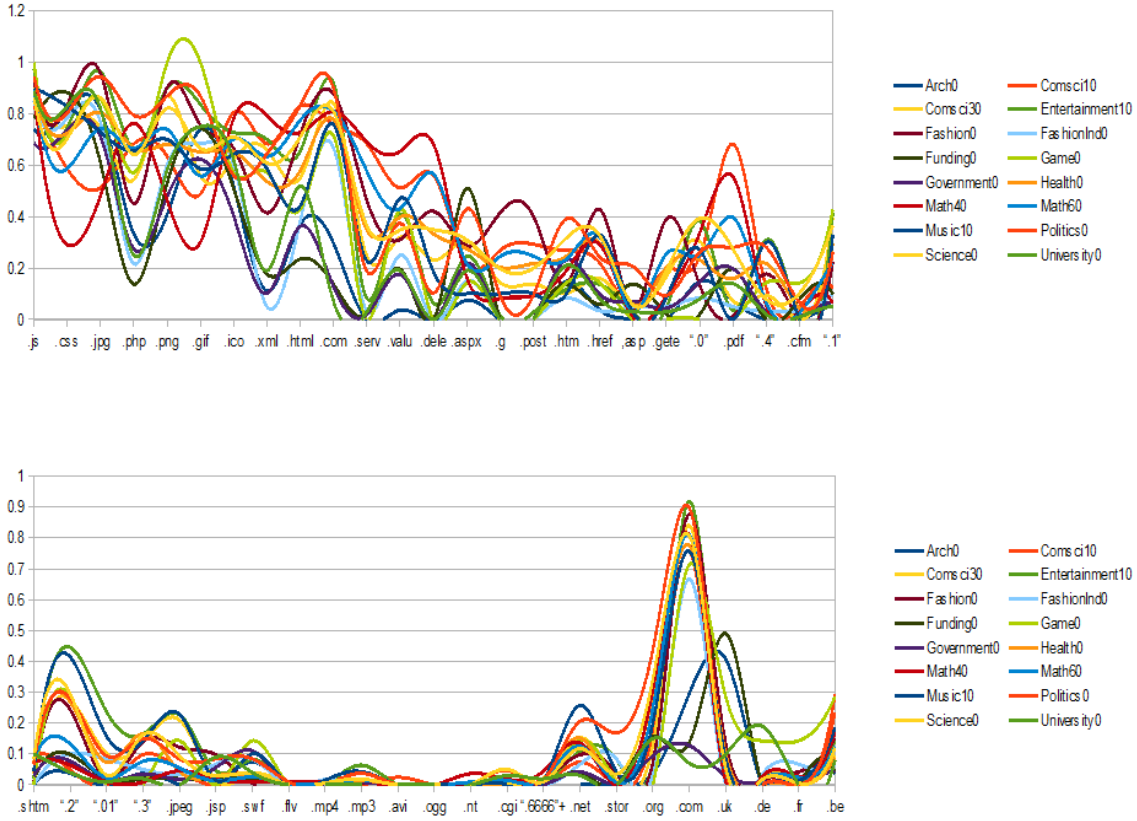


Figure 4.3-13 Ratio of URLs (y-axis) using selected patterns (x-axis) across sixteen subcategories (represented as different colour lines). The patterns statistics have been divided into two images in order to make the details more visible.

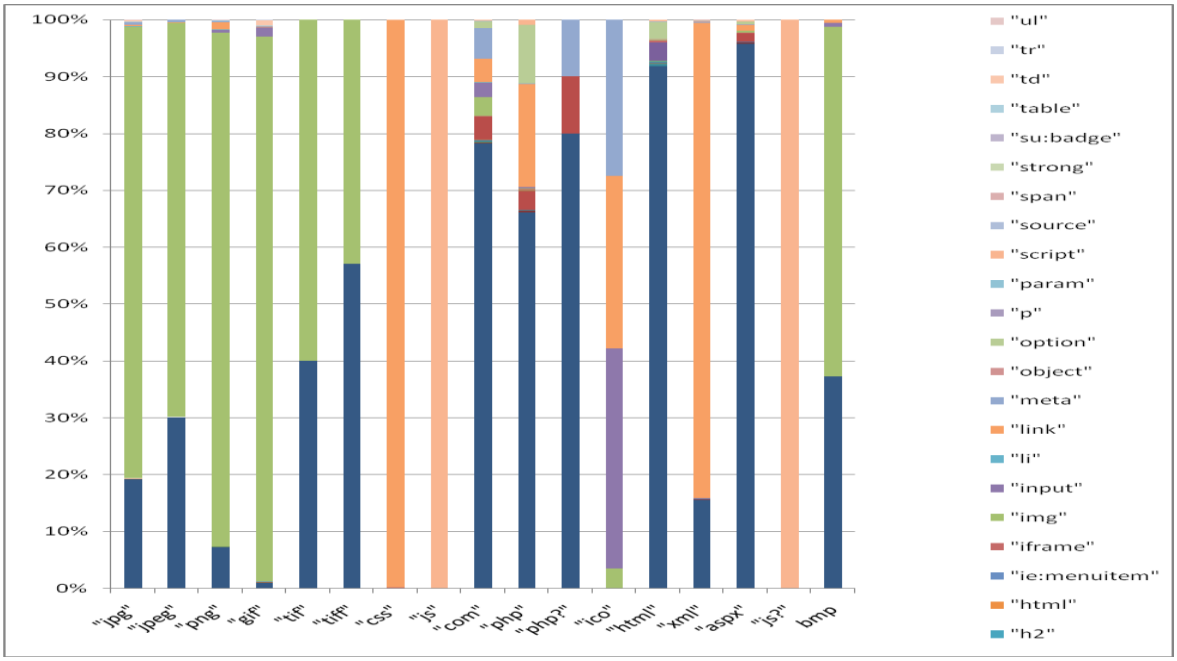


Figure 4.3-14 Percentage of tags (represented by colours along the columns) with respect to common file extensions (x-axis).

The patterns we discussed so far, on their own do not convey a sense of what type of resources these patterns represent. However, when these patterns are associated to their corresponding HTML tags, their use context can be often determined. In Figure 4.3-14, we have visualised the relationship between selected patterns and the HTML tags from whose attribute field value we have extracted the patterns. From this we can immediately associate, for example, patterns like ".js" to a type of script, and ".jpg" to a type of image and ".css" to linked resources.

Of course, these patterns are already well known to us, but within the context of automated object identification, format identification, and long term preservation, the patterns may prove useful.

4.3.6 Networking Structure

The analysis of network structures are a extremely involved process and there is much controversy surrounding the best measures and key patterns to study in a network structure (Coulon 2005). The study presented here is intended to be only a preliminary step toward highlighting the potential of network analysis for characterising a blogging community and how this relates to digital object types.

The network structure that emerges from a dataset can vary considerably. As an example, in Figure 4.3-15, we have included the network structures of six of the collection in our Categorised dataset. Each node in the structure represent a webpage in the collection and the edges represent instances where a page has provided a hyperlink to the other blogs using the "href" attribute field value for the HTML tag <*a*>. The representation here is based on 100 random pages from each of the collections (except for FashionInd0, where only 60 pages were available). The self-reference is not shown. The figure clearly shows significantly more edges in the Math40 network than in any of the other networks, In particular, there is noticeable distinction between blogs and non-blogs with respect to the number of connections arising.

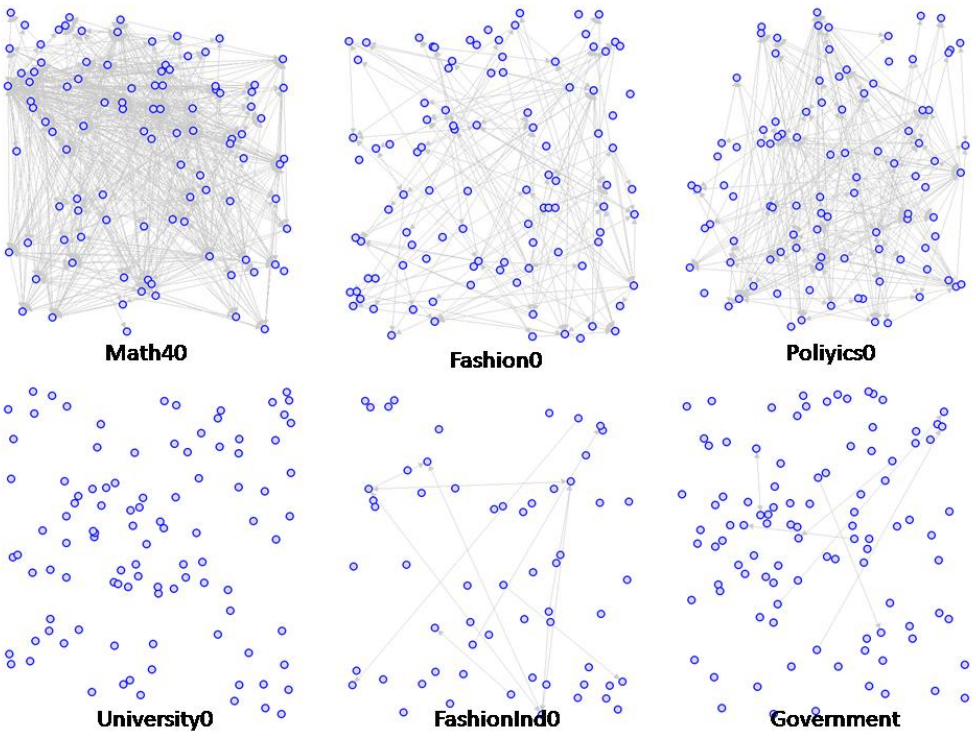


Figure 4.3-15 Network structures for three blog collections (top row) and three non-blog collections (bottom row).

The average numbers of links per page across the sixteen subcategories of the Categorised dataset are displayed in Figure 4.3-16. We again observed that the average numbers of links per page with respect to non-blogs are much smaller than those for blogs.

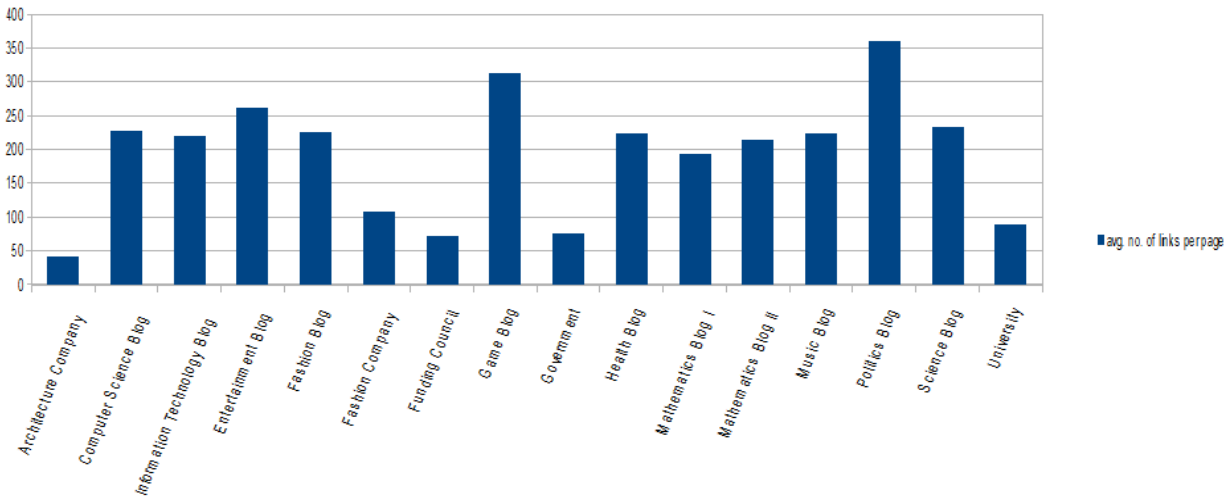


Figure 4.3-16 Average Number of hyperlinks (y-axis) found in a page: examined across the sixteen subcategories (x-axis) of Categorised dataset.

The distinction is not as pronounced when the same numbers are shown for the larger datasets (see the Mokono and Spinn3r09 average compared against the ClueWeb09 average in Figure 4.3-17). However, further investigation shows that the distinction between blogs and non-blogs lies is their tendency to self-reference. By "self-reference", we mean a hyperlink that point to a location that is a subdirectory of the associated blog.

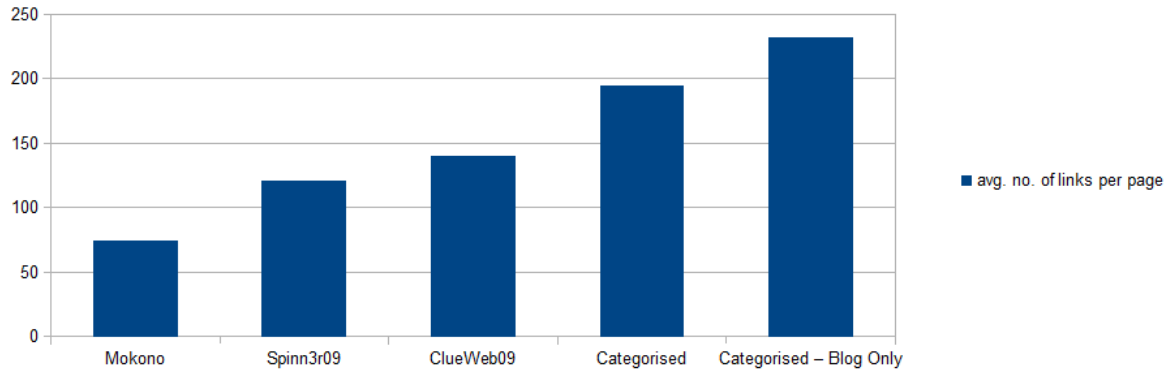


Figure 4.3-17 Average number of hyperlinks (y-axis) for the four large dataset (x-axis). The right most column is the figure for the subset of the Categorised dataset represented by blogs.

In Figure 4.3-18, we present the relative number of self-references against the number of all hyperlinks, across the datasets. The results show that the datasets which consist, predominantly, of blogs contain less number of self-references.

The difference is more noticeable when we examine the numbers with respect to the subcategories of the Categorised dataset (Table 4.3-9). The numbers in the table show that, on average, more than 76% of hyperlinks in non-blogs are self-referential. The table shows some variation across blog categories as well. Genres such as Game, Entertainment and Information Technology seem to be more self-referential than most. Mathematics Blogs also seem to contain a fair number of self-references as well. Surprisingly, the lowest rate of self-reference is associated with Fashion Blogs.

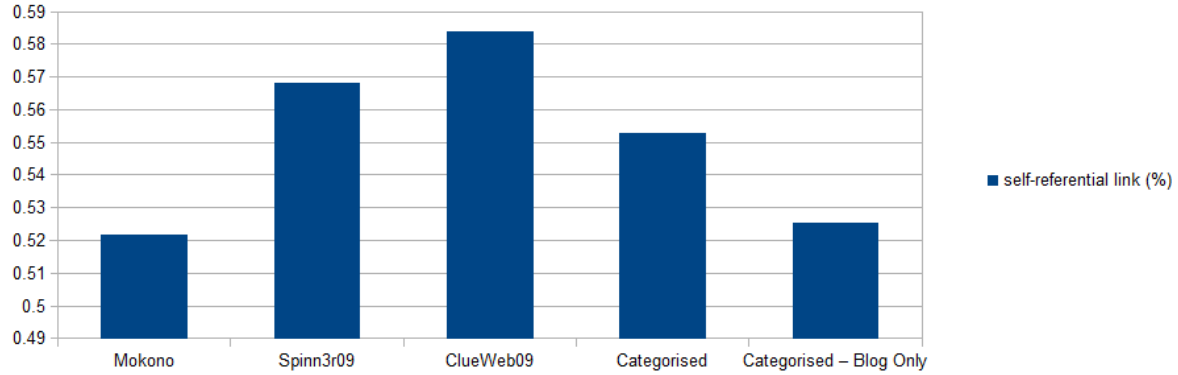


Figure 4.3-18 Ratio of self-references (y-axis) across four datasets (x-axis). The right most column is the figure for the subset of the Categorised dataset represented by blogs.

Table 4.3-9 Hyperlinks, distinct hyperlinks, and self-references in webpages. Rows in green highlight non-blog collections.

Subcategory	All Hyperlinks	Distinct Hyperlinks	No. Non-Self Referential (repeated references allowed)	Self-Referential (%)
Architecture Company	1129	843	125	0.8892825509
Computer Science Blog	9297	6823	4773	0.4866085834
Information Technology Blog	30300	21493	10313	0.6596369637
Entertainment	28733	19357	9960	0.6533602478

Subcategory	All Hyperlinks	Distinct Hyperlinks	No. Non-Self Referential (repeated references allowed)	Self-Referential (%)
Blog				
Fashion Blog	36940	28660	23513	0.3634813211
Fashion Company	6440	4926	1154	0.8208074534
Funding Council	3664	2803	503	0.8627183406
Game Blog	2184	1479	714	0.6730769231
Government	43356	31524	8464	0.8047790387
Health Blog	29175	21408	14054	0.5182862039
Mathematics Blog I	21360	15251	8977	0.5797284644
Mathematics Blog II	118471	83283	44349	0.6256552236
Music Blog	15675	11357	8959	0.4284529506
Politics Blog	38636	27709	20163	0.4781292059
Science Blog	249652	155420	129754	0.4802605226
University	8983	7369	2095	0.7667816988

The results in this section indicate that the large number of hyperlinks is one of the most distinguishing features of a blog when compared to a non-blog. In particular, the ratio of non-self-referential blogs is much higher in blogs than in other pages. This implies that blogs like to link to other resources.

We also examined the density of networks (the number of edges in the network divided by the maximum possible number of edges) with respect to each subcategory of the Categorised dataset. We retrieved 1000 blog samples of size 100 and analysed the links that exist between them. This resulted in the highest density being associated with Mathematics. The average number of edges per node in the Mathematics network is more than 8, while that in the Politics network is around 4. This observation, together with the results in Figure 4.3-17 and Table 4.3-9, allows us to identify the behavioural features that distinguish the two domains: a) both Mathematics blogs and Politics blogs include a lot of references to resources, but, b) Mathematics blogs tend to reference their own articles and articles in other Mathematics blogs, while Politics blogs tend to reference externally to resources that are not necessarily other Politics blogs. Intuitively this makes sense: most likely mathematicians blog to form discussion groups on specific mathematical topics, while Politics will be driven by events that take place in society.

4.3.7 User Generated Categories and Tags

User generated tags are characteristic of social networking media technology. Consequently, the generation of topic tags is expected to be fairly active in the blogging communities. As expected there were no user generated categories and/or tags in the non-blog subcategory collections of the Categorised dataset except a set of six categories ("tv", "tourism", "sports-news", "somali-politics", "somali-news", and "business-news") that were used in one of the government home pages. The page did not use any further topic tags. Surprisingly, we could not extract any categories or topic tags from the Mokono dataset. Either the platform at Mokono does not use the pattern "*category-term*" and "*tag-term"* for their user generated categories/tags or the bloggers do not use any.

The category and tag statistics, across the three large datasets and blogs in Mathematics Blog I, are presented in Figure 4.3-19 and Figure 4.3-20. The figures clearly show that, among these four datasets, Mathematics Blog I is the most active in sharing information. Whether the lack of shared

categories can be considered to be a lack of interaction between the blogs need to be investigated further.

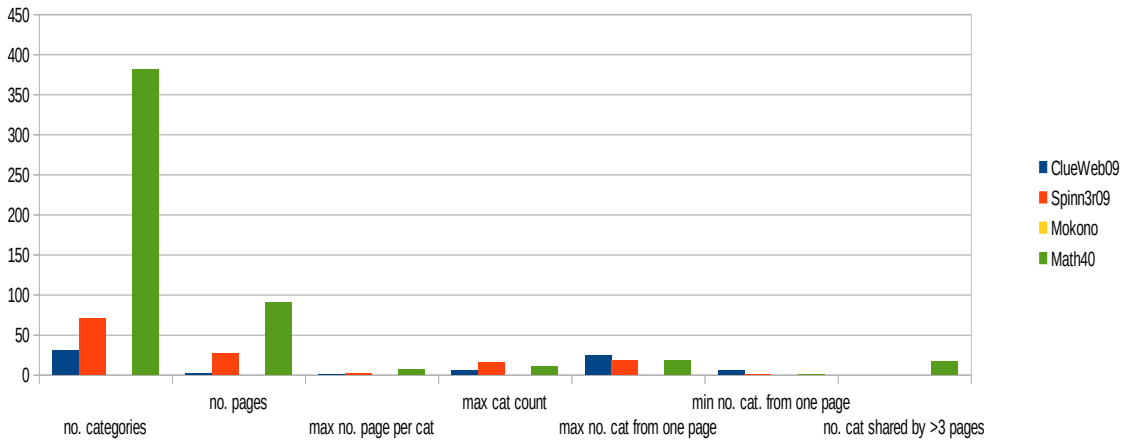


Figure 4.3-19 User generated category statistics across four datasets (represented in colour): the number of categories extracted, the number of pages that provided them, maximum number of pages found sharing one category, maximum number of times a single category is used, maximum number of categories used in any one page, minimum number of categories used in any one page, number of categories shared by at least three pages.

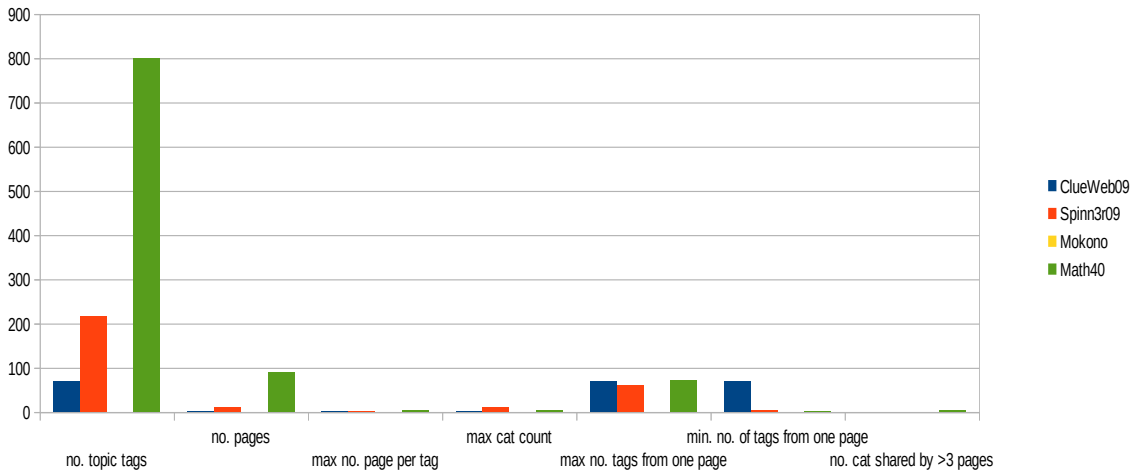


Figure 4.3-20 User generated topic tag statistics across four datasets (represented in colour): the number of tags extracted, the number of pages that provided them, maximum number of pages found sharing a tag, maximum number of times a tag is used, maximum number of tags used in any one page, minimum number of tags used in any one page, number of tags shared by at least three pages.

The results in this section suggest evidence that the user generated categories and topic tags might be useful in determining the level of information sharing activity taking place in the blogs. In digital preservation, especially in more recent topic areas involving complex objects, it has become clear that traditional approaches to metadata description do not capture interactive elements associated to the digital information. The study here could not be taken forward due to the lack of time and resources of the project, however, the investigation of information sharing behaviour should be carried forward as relevant research in the area of weblog reservation.

4.4 Conclusions

In this chapter we discussed how we might conduct preservation strategy testing within the context of weblog preservation. We discussed the problems that under lie preservation testing, especially in the context of weblogs. We reviewed the previous approaches to strategy testing (Section 4.1). We discussed risks of information loss and how the identified risks raise questions of complexities to the weblog preservation problem (Section 4.2). These complexities could render preservation processes (such as emulation, migration, normalisation, and standardisation) unscalable (Section 4.2.3), and, in some cases introduce risks of information loss themselves (Section 4.2.2).

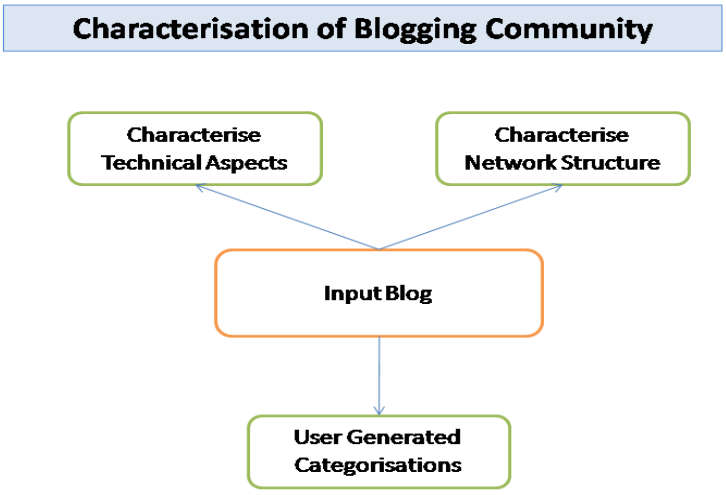


Figure 4.4-1 The three aspects of measuring weblog complexity.

As a solution to the problem we propose an approach based on measuring weblog complexity. The approach is based on the characterisation of a blog by profiling three aspect of the weblog (see Figure 4.4-1):

1. Technical characteristics based on HTML declarations, HTML tags, attribute fields, and attribute field values (Figure 4.4-2):
    - The variety of declarations, tags and attribute values, as well as their relative frequency of usage can indicate the level of syntactic complexity that the repository will need to handle.
    - The same profile can provide the scope of object types and formats that are likely to need preservation and management support.
    - The profile can serve as a robust measure to bench-mark datasets for testing preservation strategies.
2. Network characteristics based on the number of hyperlinks, the number of self-referential hyperlinks, the network density and centralities (Figure 4.4-3):
    - These measures inform us about the semantic dependencies that the weblog might have on external and internal resources. This supports identifying risks of information loss.
    - This can be combined with the recommendation/request repository features to improve the quality of the repository (Figure 7.3-1, Chapter 7).
    - These measures might provide us with a means of implementing an automated selection process.

3. Information sharing characteristics based on the variety of user generated categories and topic tags, and the number of categories and tags shared with other blogs or networks (Figure 4.4-4):
    - This will give an understanding of different topic areas that need to be supported.
    - It could serve to characterise the designated community with respect to the preservation of the blog.
    - Helps to capture the interactive properties of the blog.

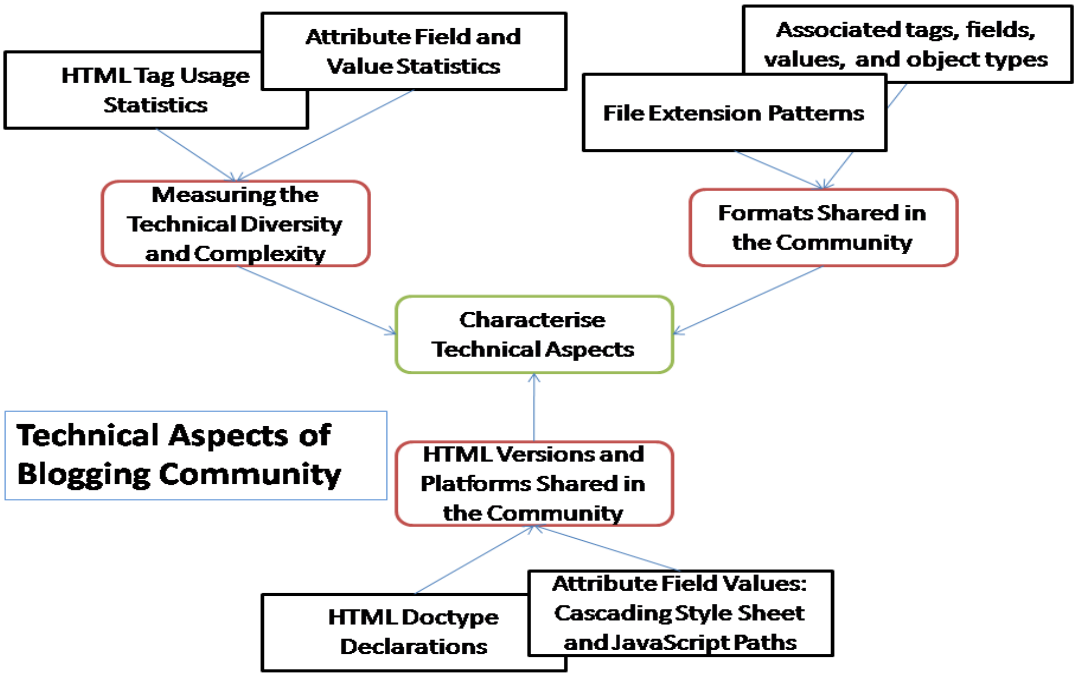


Figure 4.4-2 Identifying technical characteristics of a weblog.

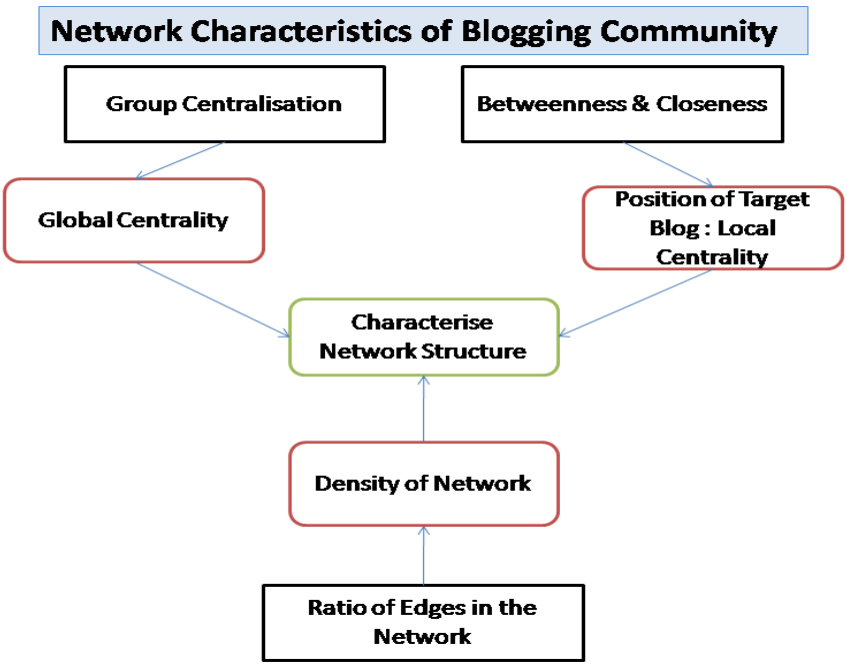


Figure 4.4-3 Identifying Network characteristics of the weblog.

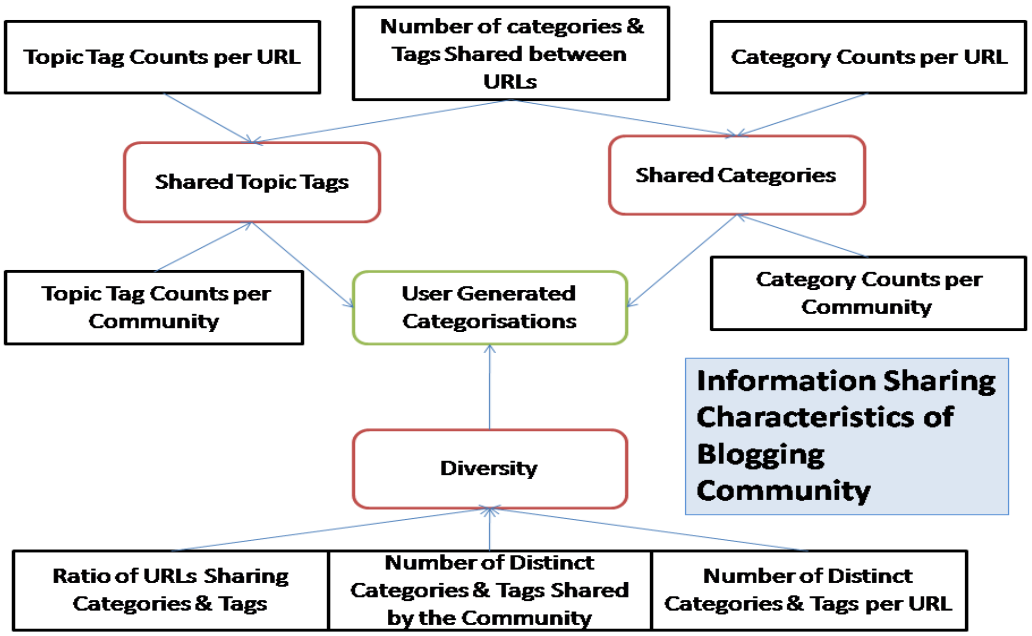


Figure 4.4-4 Identifying information sharing characteristics of a weblog.

In Section 4.3, we provided evidence that the proposed profiles are useful in diagnosing the complexities of a weblog, characterising the source community by which the weblog has been produced. For example, we have shown that all three aspects are useful in distinguishing between a blog and a non-blog page (e.g. using attribute field value patterns such as ".valu", ".serv"; analysis of self-referential hyperlinks that reveal differences across domains; the existence or non-existence of user generated categories), between blogs from different communities (distinct sets of user generated categories and different levels of sharing behaviour) , and understand the pragmatic context (e.g. the distinguishing practices of bloggers in Mathematics Blog and Politics Blog – discussed at the end of Section 4.3.6).

The characterisation of weblog complexity as outlined here, if exposed and shared, will:

- help us select datasets representative of different complexities in carrying out preservation strategy/process testing and use case studies;
- support preservation planning by mapping some of the predictable challenges in advance,
- serve as the first step toward enabling the digital curation community to compare preservation strategy tests across the board;
- help us to develop an automated approach to deriving significant properties of digital information (the characterisation described here is not inherently specific to weblog profiling), and;
- function as a digital finger print of community that produced the digital information, a valuable trace of our technological history, and supporting evidence for determining authenticity of information.

The concluding proposal is that a characterisation of blogging community similar to the one described here (Figure 4.4-1) should form part of the object characterisation and metadata assignment stage of the repository ingest phase (to be described in Section 6.1.2 and refined in Section 7.2 (to include the community characterisation process described here).

5 Recommended Metadata Schemas

The weblog survey (Section 3.1), weblog data model and resulting repository record types (Section 3.2), user requirements (Section 3.3), and significant properties of selected object types (Section 3.4) described in the last chapter indicates the types of information that we will need to be gathering to describe, manage and use the repository holdings. In this chapter, we review and recommend a range of metadata schemas and standards for these materials. Our aim is to identify a set of metadata standards that will help to ensure the authenticity, integrity, completeness, usability, and long term accessibility of the preserved content. To this end we wanted to identify metadata that met a number of robust criteria.

Metadata is structured data which describe the characteristics of a resource. Metadata are commonly defined as data about data. A metadata record consists of a number of pre-defined elements representing specific attributes of a resource, and each element can have one or more values.

A metadata schema will usually have the following characteristics:

- a limited number of elements
- the name of each element
- the meaning of each element

Metadata is an essential part of any digital resource. If a resource is to be retrieved and understood and maintained over time it must be described in a consistent, structured manner suitable for processing by computer software. Access to digital information over time is at risk unless we have good metadata standards covering the relevant defined tasks. [176]

5.1 Criteria for Selecting Metadata Schema

The metadata schemas were reviewed on the basis of the following ten criteria, devised by the University of London using a methodology adapted from existing published sources:

1. Fit for purpose
2. Open / non-proprietary standard
3. Ubiquitous / widely adopted and used / implemented
4. Has a maintenance agency or good support community
5. Well-documented /good quality documentation
6. Interoperable
7. Format-specific and covers all formats in scope
8. XML based
9. Integrates with METS
10. Integrates with PREMIS

The criteria were chosen to meet the explicit requirements outlined below:

#1. This criterion is intended to ensure that the selected metadata standard would in fact support the creation and management of technical metadata, which is the purpose underlying the task.

#2-7. These criteria were influenced by The National Archives (TNA) document *Selecting File Formats for long-term preservation* (http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf, Adrian Brown, 2008).

[176] See for example *Choosing a Metadata Standard for Resource Discovery: A QA Focus Document*, UKOLN Briefing Paper #63 (2006)

Obviously this guidance note is concerned with selecting file formats, but the advice is sound and that the basic principles here are a good fit for metadata standards also, especially as the report is a survey of available metadata standards which could apply to the principal digital object types already identified.

#8. This criterion was selected because XML is understood to be the easiest way to implement PREMIS and METS. It is also explicitly stated in the project Description of Work (DoW) that XML is "the vehicle to convey the metadata necessary for both the management of digital objects within a repository but also the exchange of such objects between repositories."

#9-10 These criteria were selected because the project DoW has already explicitly stated that PREMIS and METS are in scope for evaluation. See Part B pp 14, 18, and 31.

The terminology is explained:

1. **Fit for purpose**: the metadata standard must be fit for managing technical metadata. Technical Metadata expresses technical details of the stored resource necessary to identify, validate and preserve the content. Technical metadata is information regarding files' creation, format, and use characteristics[177].

2. **Open / non-proprietary standard**: the standard selected must be open and non-proprietary to avoid lock-in to an industry standard.

3. **Ubiquitous / widely adopted and used / implemented**: the standards considered must be widely used and implemented as this will enable interoperability, crosswalks and is indicative of an active community of support. This will ensure support for the standard over time and provide assurances regarding preservation.

4. **Has a maintenance agency or good support community**: the standard must have an active community supporting it. Using a standard that is well-supported can also bring cost benefits. Implementation guidance, user guidance, examples, XML/RDF schemas, crosswalks, multi-lingual capacity, and software tools may pre-exist, thus easing the process of development, customisation and update.

5. **Well-documented / good quality documentation**: the metadata standard must be well-documented with the documentation made freely available.

6. **Interoperable**: the metadata standard must be selected from a leading standard within the community or domain. This will help to make the resource accessible beyond the confines of the project. Metadata that is in a recognisable common schema may be harvested by subject or domain-wide portals and cross-searched with resources from many other institutions. This is unlikely to happen if an in-house/non interoperable standard is used.

7. **Format-specific and covers all formats in scope**: Ensure that the metadata standards are capable of being format-specific, and that they can apply to the types of file formats identified by University of London from the blog sample in WP2: i.e. structured text, image, audio, video, media (or documents).

8. **XML-based**: The metadata standard ought to be XML-based, or at least expressible in the form of an XML schema. [178] XML is a standard which works well for document definition and defining

[177] http://www.loc.gov/standards/mets/METSOverview.v2.html
[178] http://www.w3.org/XML/Schema

structure and representation. XML can be kept in a human-readable format, and users can understand and edit it without specialized tools. This requirement overlaps with 9 and 10.

9. **Integrates with METS**: The selected standards ought to integrate or be compatible with METS [179]; the use of METS profiles has been explicitly declared as a BlogForever project aim. Any number or type of digital files can be described and linked together by a METS record, enabling it to represent very complex digital resources. METS can embed or link to many XML-based metadata (e.g. MODS, MIX, PREMIS or TEI). METS can be understood as a binder/wrapper that unites descriptive and technical metadata about a particular resource. A METS record includes six parts: Header, Descriptive metadata, Administrative metadata, File groups, Structural map, and Behaviour section.

10. Integrates with PREMIS: By the same rationale, the selected standards ought to integrate with the PREMIS (Preservation Metadata: Implementation Strategy) standard [180.] PREMIS consists of a core set of standardized data elements that are recommended for repositories to manage and perform the preservation function. These crucial functions include actions to make the digital objects useable over time, keeping them viable, or readable, displayable and kept intact, all for the purpose of future access. Additionally, a PREMIS schema can be wrapped up in a METS profile.

5.2 Deciding whether a Schema Meets the Criteria

The guidelines below were used to evaluate a given standard as meeting the criteria:

The values in the tables were justified by matching them against aspects of the METS standards in each case. The fact that METS enjoys Library of Congress (LOC) support means the standards meet most of the criteria. To put it another way, if it's an LOC METS schema, it is already de facto a good match for #2, 4, 5, 6, 8, 9, and 10. See detail of these statements below.

If evidence was found of other repositories using the schema in question, it ticks box #3. If it's specific to a digital object type, as already established, then it's a #7. If it scores all the other nine boxes, then it's fair to say it's fit for purpose, i.e. #1.

1. Fit for purpose: "The METS schema is a standard for encoding descriptive, administrative, and structural metadata"
2. Open / non-proprietary standard. "Any METS document has the following features: An **open standard (non-proprietary**)" (http://en.wikipedia.org/wiki/METS)
3. Ubiquitous / widely adopted and used / implemented: see http://www.loc.gov/standards/mets/mets-examples.html for examples by numerous METS implementers, and the METS implementation registry for descriptions of METS projects planned, in progress, and fully implemented. http://www.loc.gov/standards/mets/mets-registry.html
4. Has a maintenance agency or good support community: "The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress"
5. Well-documented /good quality documentation: see http://www.loc.gov/standards/mets/mets-schemadocs.html for examples of documentation.
6. Format-specific and covers all formats in scope (i.e. is specific to a digital object type).
7. XML based: "The METS schema is expressed using the XML schema language"
8. Integrates with PREMIS. See for example http://old.diglib.org/forums/spring2008/presentations/Habing.pdf, http://www.loc.gov/standards/mets/presentations/Olaf_Brandt_PREMIS_and_METS.pdf,

[179] http://www.loc.gov/standards/mets/
[180] http://www.loc.gov/standards/premis/index.html

http://www.dlib.org/dlib/september08/dappert/09dappert.html,
http://ddp.nist.gov/workshop/papers/03_03_nist-rguenther-premis.pdf

5.3 Descriptive Metadata Schema

Descriptive Metadata Section of METS <dmdSec> Contains descriptive metadata, supplying information on the intellectual content of an object which is necessary for users to find an item and assess its value for their research. It may contain the metadata itself, or point to metadata held outside the METS document. Multiple instances of both external and internal descriptive metadata may be included. For external metadata the <mdRef> element allows the provision of a URI for that metadata.[181]

As a first step we went through the list of generally available metadata standards for all possible types of objects[182]. From this list we selected the metadata standards that can be potentially used to describe complex digital objects, a definition which includes blogs and blog posts. Finally, from those we selected the ones that focus in describing the digital objects themselves, based on the definition of descriptive metadata within METS, as mentioned above.

There are three available descriptive metadata standards that are potentially useful for BlogForever: MARCXML[183], Dublin Core[184] and MODS[185].

5.3.1 MARCXML

MARC XML is an XML schema based on the fairly common MARC21 standard. MARC (MAchine-Readable Cataloging) is a data format and set of related standards used by libraries to encode and share information about books and other material they collect. It was first developed by Henriette Avram at the Library of Congress in the 1960s, and is still widely used today as the basis for most online public access catalogs. MARCXML was developed by the US Library of Congress and adopted by it and others as a means of easy sharing of, and networked access to, bibliographic information. Being easy to parse by various systems allows it to be used as an aggregation format. The MARC XML primary design goals included:

- Simplicity of the schema.
- Flexibility and extensibility.
- Lossless and reversible conversion from MARC.
- Data presentation through XML stylesheets.
- MARC records updates and data conversions through XML transformations.
- Existence of validation tools.

One of the MARC formats is the Bibliographic records. They describe the intellectual and physical characteristics of bibliographic resources (books, sound recordings, video recordings, and so forth).

5.3.2 Dublin Core

The Dublin Core set of metadata elements provide a small and fundamental group of text elements through which most resources can be described and catalogued. Using only 15 base text fields, a Dublin Core metadata record can describe physical resources such as books, digital materials such as video, sound, image, or text files, and composite media like Web pages. Metadata records based on Dublin Core are intended to be used for cross-domain information resource description and have

[181] http://www.paradigm.ac.uk/workbook/metadata/mets-structure.html
[182] http://en.wikipedia.org/wiki/Metadata_standards
[183] http://www.loc.gov/standards/marcxml/
[184] http://dublincore.org/metadata-basics/
[185] http://www.loc.gov/standards/mods/

become standard in the fields of library science and computer science. Implementations of Dublin Core typically make use of XML and are Resource Description Framework based.

The Simple Dublin Core Metadata Element Set (DCMES) consists of 15 metadata elements: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights.

5.3.3 MODS

The United States Library of Congress' Network Development and MARC Standards Office, with interested experts, developed the Metadata Object Description Schema (MODS) in 2002 for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications. As an XML schema it is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format.

MODS was designed to be more end-user oriented than the full MARCXML schema, and its element set is simpler than the full MARC format. However an original MARC 21 record converted to MODS may not convert back to MARC 21 in its entirety without some loss of specificity in tagging or loss of data.

Our sense is that MODS has not yet been widely adopted as a metadata schema, although we note that the Library of Congress have used it to catalogue their web archive collection.186 This implementation has potential for BlogForever, but it appears to be providing catalogue access at a fairly limited level and thus may not offer enough richness of detail for describing blog content.

5.3.4 Comparison Against Criteria

In order for the ten criteria listed at the beginning of this chapter to apply to the case of “Descriptive metadata” and be compatible with the existing repository framework for BlogForever (Invenio[187]), the “Integrates with PREMIS” criterion has been replaced with “Integrates with Invenio”.

MARCXML

Table 5.1-1: compatibility of MARCXML

Criterion	Met	Criterion	Met
Open source	YES	Integrates with METS	YES
Widely adopted	YES	Covers formats in scope	YES
Maintained / supported	YES	XML-based	YES
Documented	YES	Integrates with Invenio	YES
Interoperable	YES		

Dublin Core

Table 5.1-2: compatibility of Dublin Core

Criterion	Met	Criterion	Met
Open source	YES	Integrates with METS	YES

[186]See http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

[187] http://invenio-software.org/

Widely adopted	YES	Covers formats in scope	YES
Maintained / supported	YES	XML-based	YES
Documented	YES	Integrates with Invenio	NO
Interoperable	YES		

MODS

Table 5.1-3: compatibility of MODS

Criterion	Met	Criterion	Met
Open source	YES	Integrates with METS	YES
Widely adopted	NO (not as much, no clear data)	Covers formats in scope	YES
Maintained / supported	YES	XML-based	YES
Documented	YES	Integrates with Invenio	NO
Interoperable	YES		

All three schemas go some way to meeting these criteria, but MARC is already integrated in Invenio as the underlying bibliographic standard of the system. MARCXML supports interoperability with other digital libraries. Additionally, the element set is richer than Dublin Core.

5.3.5 Example of MARC in METS

Assuming METS is selected for implementation in BlogForever, there are two approaches to express MARC inside a METS wrapper:

1. **Internal Descriptive Metadata (mdWrap)**: An mdWrap element provides a wrapper around metadata embedded within a METS document. Such metadata can be in one of two forms:
    a. XML-encoded metadata, with the XML-encoding identifying itself as belonging to a namespace other than the METS document namespace.

```
<dmdSec                                                    ID="dmd002">
    <mdWrap MIMETYPE="text/xml" MDTYPE="MARC" LABEL="MARC Metadata">
        <xmlData>
            (your                    MARCXML                    here)
        </xmlData>
    </mdWrap>
</dmdSec>
```

    b. A binary or textual form, provided that the metadata is Base64 encoded and wrapped in a element within the mdWrap element.

```
<dmdSec                                                    ID="dmd003">
    <mdWrap  MIMETYPE="application/marc"  MDTYPE="MARC"  LABEL="OPAC
Record">
        <binData>MDI0ODdjam0gIDIyMDA1ODkgYSA0NU0wMDAxMDA...(etc.)
        </binData>
    </mdWrap>
</dmdSec>
```

2. External Descriptive Metadata (mdRef): an mdRef element provides a URI which may be used in retrieving the external metadata. For example, the following metadata reference points to the finding aid for a particular object:

```
    <mets:dmdSec                                                    ID="dmd004">
        <mets:mdRef xlink:href="http://roger.ucsd.edu/record=b3904109"
                    LOCTYPE="URL"
                    MDTYPE="MARC"
                    LABEL="Some Record"/>
</mets:dmdSec>
```

5.3.6 Example of Blogs in MARC

After choosing one of the previous options to embed MARC in METS, we need to define how to use MARC to describe blogs. The next table is a draft mapping between the conceptual data model from BlogForever: D2.2 Weblog Data Model and MARC tags. It is just one possible example solution, since the final implementation is to be decided in a later stage of the design.

Some of the attributes proposed in D2.2 have been matched because we understand that their meaning is similar enough, if not the same. Again, this is just a draft used for the example. The complete description and meaning of each tag are described by the Library of Congress.[188]

Table 5.1-4: mapping of data model to MARC tags

post	Comment	MARC tag	comments
title	Subject	245 $a	
subtitle		245 $b	
URI	URI	520 $u	Uniform Resource Identifier
date_created	date_added	269 $c	Used in Invenio
date_modified	date_modified	005	Date and Time of Latest Transaction
version		075	NEW
status_code	Status	076	NEW
geo_longitude	geo_longitude	342 $g	Geospatial Reference Data, longitude
geo_latitude	geo_latitude	342 $h	Geospatial Reference Data, latitude
visibility		506	Restrictions on Access Note
has_reply	has_reply	788 $a	NEW
last_reply_date		788 $c	NEW
num_of_replies	num_of_replies	788 $b	NEW
child_of	is_child_of_post is_child_of_comment	773 $w $4	We can use several tags, one for the blog it belongs to, another for the post it belongs to (for comments only), another for the comment it is

[188] http://www.loc.gov/marc/bibliographic/

post	Comment	MARC tag	comments
			replying to (only for comments on comments)
type		336	Content Type
posted_via		781	NEW
addressed_to_URI	addressed_to_URI	789	NEW
previous_URI		780	Preceding Entry
next_URI		785	Succeeding Entry
full_content	full_content	520 $a	Currently used for 'abstract' in Invenio
full_content_format	full_content_format	520 $b	the escaped html code
note	Note	500 $a	
encoding	Encoding	532	NEW
copyright	Copyright	017 $*	

Following this table, this would an example of a blog post described with MARC. We have chosen a post in the BlogForever project blog. The HTML version of the content has been escaped to avoid conflicts between MARCXML tags and HTML tags.

```
<collection xmlns="http://www.loc.gov/MARC21/slim">
  <record>
    <controlfield tag="001">0000002</controlfield>
    <controlfield tag="005">2410201117240000</controlfield>
    <datafield tag="017"></datafield>
    <datafield tag="245" ind1=" " ind2=" ">
    <subfield code="a">What relationships among blogs do you know?</subfield>
    </datafield>
    <datafield tag="520" ind1=" " ind2=" ">
    <subfield      code="u">http://blogforever.eu/blog/2011/08/08/what-relationships-
among-blogs-do-you-know</subfield>
    </datafield>
    <datafield tag="269">
      <subfield code="c">24/10/2011</subfield>
    </datafield>
    <datafield tag="075">
      <subfield code="a">1</subfield>
    </datafield>
    <datafield tag="076">
      <subfield code="a"></subfield>
    </datafield>
    <datafield tag="342">
      <subfield code="g">46.198392</subfield>
      <subfield code="h">6.142296</subfield>
    </datafield>
    <datafield tag="506">
<subfield code="a">Access copy available to the general public.</subfield>
      <subfield code="f">Unrestricted</subfield>
    </datafield>
    <datafield tag="788">
```

```
      <subfield code="a">No</subfield>
      <subfield code="b">Null</subfield>
      <subfield code="c">01</subfield>
    </datafield>
    <datafield tag="773" ind1=" " ind2=" ">
      <subfield code="w">0000001</subfield>
      <subfield code="4">blog</subfield>
    </datafield>
    <datafield tag="336" ind1=" " ind2=" ">
      <subfield code="a">post</subfield>
    </datafield>
    <datafield tag="781">
      <subfield code="a"></subfield>
    </datafield>
    <datafield tag="789">
      <subfield code="a">None</subfield>
    </datafield>
    <datafield tag="780" ind1=" " ind2=" ">
      <subfield       code="a">http://blogforever.eu/blog/2011/07/11/the-blogforever-
survey-is-live/</subfield>
    </datafield>
    <datafield tag="785" ind1=" " ind2=" ">
      <subfield       code="a">http://blogforever.eu/blog/2011/09/16/2nd-blogforever-
consortium-meeting/</subfield>
    </datafield>
    <datafield tag="520" ind1=" " ind2=" ">
      <subfield code="a">One aim in BlogForever is the analysis of the network of
blogs. Analyses of relationships between blogs can help to achieve

a better understanding of structures and processes in the blogosphere,
rankings based on network criteria,
insights in the lifecycle of blogs and blog communities, and
useful suggestions for blog authors regarding potential connections to other blogs.
Further purposes are conceivable.

An important prerequisite for network analysis is the identification of potential
relations, especially such relations that can be captured by a software agent.
Therefore, we attempt to describe the possible relationships among blogs and
bloggers. The obvious relationships that are referenced in the literature are:

A citation or link: A blog post can contain a hyperlink to another blog or blog post.
The blogroll.
Trackback or pingback functionalities.
Comments: The comment relation represents a relationship between the person who
creates a comment and the blog (or blog author) where the comment occurs.
Nevertheless, there are many more relationships, especially if you take relationships
between blogs and other media into consideration, e.g. delicious' bookmarks,
facebooks like button, etc..

But what relationships do you know? We will be happy if you tell us your proposals
for additional relationships among blogs and bloggers!</subfield>
      <subfield code="b">&lt;div class="post" id="post-733"&gt;
&lt;div class="author-box"&gt;
&lt;img                                                src="http://blogforever.eu/wp-
content/uploads/avatars/1/815cf9f0cb91ff8810647c9538dec2c2-bpthumb.jpg"        alt=""
class="avatar user-1-avatar" width='50' height='50' /&gt;
                    &lt;p&gt;by   &lt;a   href="http://blogforever.eu/members/admin/"
title="Vangelis Banos"&gt;Vangelis Banos&lt;/a&gt;&lt;/p&gt;
&lt;/div&gt;
&lt;div class="post-content"&gt;
&lt;h2   class="posttitle"&gt;&lt;a   href="http://blogforever.eu/blog/2011/09/16/2nd-
blogforever-consortium-meeting/"   rel="bookmark"   title="Permanent   Link   to   2nd
BlogForever       Consortium       Meeting"&gt;2nd       BlogForever       Consortium
Meeting&lt;/a&gt;&lt;/h2&gt;
&lt;p          class="date"&gt;7:48          pm          &lt;em&gt;in          &lt;a
href="http://blogforever.eu/blog/category/blog/"  title="View  all  posts  in  Blog"
rel="category                      tag"&gt;Blog&lt;/a&gt;,                      &lt;a
```

```
href="http://blogforever.eu/blog/category/news/" title="View all posts in News"
rel="category tag"&gt;News&lt;/a&gt; by &lt;a
href="http://blogforever.eu/members/admin/" title="Vangelis Banos"&gt;Vangelis
Banos&lt;/a&gt;&lt;/em&gt;&lt;/p&gt;
&lt;div class="entry"&gt;
&lt;p style="text-align: justify;"&gt;&lt;a href="http://blogforever.eu/wp-
content/uploads/2011/09/PYRGOS.jpg" &gt;&lt;img class="alignleft size-medium wp-
image-734" title="Thessaloniki" src="http://blogforever.eu/wp-
content/uploads/2011/09/PYRGOS-300x191.jpg" alt="PYRGOS 300x191 2nd BlogForever
Consortium Meeting" width="300" height="191" /&gt;&lt;/a&gt;The 2nd BlogForever
Consortium Meeting took place during 8-9 September in Thessaloniki, Greece. Nineteen
participants from twelve institutions came to Thessaloniki to discuss about
BlogForever. Current progress was evaluated and the project roadmap was laid
down.&lt;/p&gt;
&lt;p style="text-align: justify;"&gt;The meeting was organized in sessions covering
all aspects of the project:&lt;/p&gt;
&lt;ul&gt;
&lt;li&gt;Weblog Structure and Semantics (WP2) was one of the main sessions of the
meeting, covering recently submitted &lt;a title="The BlogForever survey is live!"
href="http://blogforever.eu/blog/2011/07/11/the-blogforever-survey-is-live/"
&gt;BlogForever Survey&lt;/a&gt; and the pending Blog Data Model.&lt;/li&gt;
&lt;li&gt;The BlogForever Policies (WP3) section of the meeting covered work on Risk
management as well as the Preservation Policy.&lt;/li&gt;
&lt;li&gt;In the BlogForever software platform (WP4) session, work on User
Requirements &amp;amp; Platform Specifications was evaluated. Additionally, a special
technical session explored possible ways of designing &amp;amp; developing the
BlogForever Platform.&lt;/li&gt;
&lt;li&gt;Last but not least, the dissemination plan &amp;amp; associated activates
were presented in the Dissemination &amp;amp; Exploitation (WP6) session.&lt;/li&gt;
&lt;/ul&gt;
&lt;p&gt;Besides BlogForever partners, &lt;a
href="http://www.mcgill.ca/sis/people/faculty/hank/"
onclick="javascript:_gaq.push(['_trackEvent','outbound-
article','http://www.mcgill.ca']);"&gt;Carolyn Hank&lt;/a&gt; was also invited to
present her work on Blog Preservation and contribute to expanding the spectrum of the
project.&lt;/p&gt;
&lt;p style="text-align: center;"&gt;&lt;a href="http://blogforever.eu/wp-
content/uploads/2011/09/P1090108.jpg" &gt;&lt;img class="aligncenter size-large wp-
image-736" title="2nd BlogForever meeting partners" src="http://blogforever.eu/wp-
content/uploads/2011/09/P1090108-1024x768.jpg" alt="P1090108 1024x768 2nd BlogForever
Consortium Meeting" width="614" height="461" /&gt;&lt;/a&gt;&lt;/p&gt;
&lt;p&gt;&amp;nbsp;&lt;/p&gt;
&lt;div style="float: right; margin-left: 10px;"&gt;&lt;a
href="http://twitter.com/share?url=http://blogforever.eu/blog/2011/09/16/2nd-
blogforever-consortium-meeting/&amp;via=blogforever&amp;text=2nd BlogForever
Consortium Meeting&amp;related=:&amp;lang=en&amp;count=vertical"
onclick="javascript:_gaq.push(['_trackEvent','outbound-
article','http://twitter.com']);" class="twitter-share-
button"&gt;Tweet&lt;/a&gt;&lt;script type="text/javascript"
src="http://platform.twitter.com/widgets.js"&gt;&lt;/script&gt;&lt;/div&gt;
&lt;/div&gt;
&lt;p class="postmetadata"&gt;&lt;span class="tags"&gt;Tags: &lt;a
href="http://blogforever.eu/blog/tag/auth/" rel="tag"&gt;AUTH&lt;/a&gt;, &lt;a
href="http://blogforever.eu/blog/tag/meeting/" rel="tag"&gt;Meeting&lt;/a&gt;, &lt;a
href="http://blogforever.eu/blog/tag/thessaloniki/"
rel="tag"&gt;Thessaloniki&lt;/a&gt;&lt;br /&gt;&lt;/span&gt; &lt;span
class="comments"&gt;&lt;a href="http://blogforever.eu/blog/2011/09/16/2nd-
blogforever-consortium-meeting/#respond" title="Comment on 2nd BlogForever Consortium
Meeting"&gt;No Comments &amp;#187;&lt;/a&gt;&lt;/span&gt;&lt;/p&gt;
&lt;/div&gt;&lt;/div&gt;&lt;/div&gt;</subfield>
    </datafield>
    <datafield tag="532">
      <subfield code="a">UTF-8</subfield>
    </datafield>
    <datafield tag="017">
      <subfield code="a">Copyright information</subfield>
    </datafield>
```

```
  </record>
</collection>
```

5.4 Administrative Metadata

"Administrative" is understood to be the METS definition of the term, that is:

> Administrative Metadata Section <amdSec> of a METS profile contains technical information about the digital object, rights management information and provenance information. It is divided into four main sections: *technical metadata* (re. file creation, format and use characteristics); *IPR metadata* (re. copyright, licensing etc); *source metadata* (re. the analogue source from which a digital object derives, where relevant); *digital provenance metadata* (re. source of files, relationships between files, information about any migration or other preservation activities undertaken). [189]

The terms *technical metadata* and *digital provenance metadata* best describe what PREMIS does. *Source metadata* is not relevant to Blog Forever since we are not deriving digital objects from analogue sources. *IPR metadata* is being dealt with by another BF partner, although it is possible to manage this in PREMIS and in METS.

5.4.1 Technical Metadata

As we know there are many formal standards. This section is primarily concerned with the selection of *technical metadata*. Technical metadata associated with a digital asset is at the heart of any preservation system. Digital objects will be rendered useless over time if no information about the technical infrastructure is managed, as this knowledge can be easily lost and the technological infrastructure can become obsolete.

This section presents a survey of available metadata standards which could apply to the principal digital object types identified in Section 3.1. In each case, the report presents:

1. The name of the digital object type
2. The name of a recommended metadata standard
3. A tick-box table confirming that the standard meets our ten criteria
4. A short description of other relevant metadata standards (where appropriate), and why they were not selected

Two of the suggested standards are supported by the Library of Congress; the suggested standard for documents was developed by the Florida Digital Archive and Harvard University Library.

Digital Object Type: Structured Text

Recommended Metadata Standard

Technical Metadata for Text (TextMD) Version 2.2, 2011
TextMD is an XML Schema designed for expressing technical metadata for textual objects. It was developed at New York University; maintenance has transferred to Library of Congress. It includes format-specific technical metadata for text.

Link: http://www.loc.gov/standards/textMD/index.html

[189] See Structure of a METS file at http://www.paradigm.ac.uk/workbook/metadata/mets-structure.html

Other metadata standards for structured text

None

Digital Object Type: Image

Recommended Metadata Standard

Metadata for Images in XML Standard (MIX), Version 2.0, 2008
MIX [190]is an XML Schema designed for expressing technical metadata for digital still images. It is based on the NISO Z39.87 Data Dictionary – Technical Metadata for Digital Still Images. It can be used standalone or as an extension schema with METS/PREMIS.

Other Metadata Standards

Digital Imaging Group 35 (DIG35): Version 1.1 of their draft metadata specification for digital images appeared in April 2001 [191.] It defined a standard set of metadata for digital images in XML that could be widely implemented across multiple image file formats. It was used by Harvard HUL in 2004.

Reason for non-selection: Descriptive metadata only, not technical

Ontology for Media Resources 1.0. This document [192] defines the Ontology for Media Resources 1.0. The term "Ontology" is used in its broadest possible definition: a core vocabulary. The intent of this vocabulary is to bridge the different descriptions of media resources, and provide a core set of descriptive properties. This document defines a core set of metadata properties for media resources, along with their mappings to elements from a set of existing metadata formats. Besides that, the document presents a Semantic Web compatible implementation of the abstract ontology using RDF/OWL. The document is mostly targeted towards media resources available on the Web, as opposed to media resources that are only accessible in local repositories.

Reason for non-selection: Has some technical metadata, but not enough for the BlogForever project

Adobe and XMP: Extensible Metadata Platform (XMP) [193] is an XML-based format modelled by Adobe after W3C's RDF (Resource Description Framework) which forms the foundation of the semantic Web initiative. Adobe makes the XMP specification freely available, and offers an open-source XMP toolkit for software developers. XMP metadata travels with the file, and can be embedded in many common file formats including PDF, TIFF, and JPEG. Metadata properties are grouped in schemas. Each schema is identified by a unique namespace URI and holds an arbitrary number of properties.

Reason for non-selection: Descriptive metadata only, not technical

EXif = Exchangeable Image File Format: This is a standard for storing interchange information in image files, especially those using JPEG compression. Most digital cameras now use the EXIF format [194]. The format is part of the DCF standard created by JEITA to encourage interoperability between imaging devices.

[190] http://www.loc.gov/standards/mix/
[191] http://www.bgbm.org/tdwg/acc/Documents/DIG35-v1.1WD-010416.pdf
[192] http://www.w3.org/TR/mediaont-10 /
[193] http://www.pdflib.com/knowledge-base/xmp-metadata/
[194] http://www.exif.org/

Reason for non-selection: Relates to performing a specific industry task

IPTC is the standard developed in the 1970's by the International Press Telecommunications Council [195.] It was initially developed as a standard for exchanging information between news organizations and has evolved over time. Around 1994, Adobe Photoshop's "File Info" form enabled users to insert and edit IPTC metadata in digital image files and so it was adopted by stock photo agencies, and other publishing businesses outside of the news media.

Reason for non-selection: Relates to performing a specific industry task

Digital Object Type: Audio

Recommended Metadata Standard

AES57-2011: AES standard for audio metadata - Audio object structures for preservation and restoration

This standard began as AES-X098B Administrative and structural metadata for audio objects, a project to collect information on all metadata issues pertaining to digital audio objects and all aspects of the digital documentation of digital audio objects. This scope includes field structures to describe and provide access to the audio content contained in digital files. It includes transfer, preservation and restoration information. It is work in progress by the Audio Engineering Society SC-03-06 Working Group on Digital Library and Archive Systems.

AES is comprehensive and granular, meaning the metadata can be repurposed; both data elements and vocabularies are included; it accommodates both digital and analog formats, including those with physical carriers and those that exist as streams of bits; it has a rigorous delineation of metadata types that make it compatible with METS; it is expressed as XML and supports segmenting and long-term storage.

The standards developed in AES-X098B were released in September 2011 by the Audio Engineering Society as AES57 AES standard for audio metadata – audio object structures for preservation and restoration.

Link: http://www.aes.org/publications/standards/search.cfm?docID=84

Other Metadata Standards for Audio

AMD Schema at Library of Congress [196] ; as noted above this is going to be replaced by AES57: AES standard for audio metadata - Audio object structures for preservation and restoration.

AES60-2011 standard for core audio metadata, published in Sept 2011: [197] AES60-2011 addresses the creation, management and preservation of material that can be re-used as originally produced, or may provide input material for new production projects. Material is expected to be exchanged between various organisations or between production facilities in a distributed environment. The core set of metadata presented in this specification is a co-publication of EBU Tech3293-2008 EBU Core, itself an extension to and a refinement of the Dublin Core. EBUCore is a minimum list of attributes characterizing video and / or audio media resources. An XML representation is also provided in case this metadata would be implemented, for example in archive

[195] http://www.iptc.org/site/Home/
[196] http://www.loc.gov/standards/amdvmd/audiovideoMDschemas.html
[197] http://www.aes.org/publications/standards/search.cfm?docID=85

exchange projects using the Open Archive Initiative's Protocol for Metadata Harvesting (OAI-PMH).

Material Exchange Format (MXF) [198:] Object-based file format that wraps video, audio, and other bitstreams ("essences"), optimized for content interchange or archiving by creators and/or distributors, and intended for implementation in devices ranging from cameras and video recorders to computer systems. In effect, the format bundles the essences and what amounts to an "edit decision list" (data used by audio-visual content editing systems) in an unambiguous way that is essence-agnostic and metadata-aware. Extensive metadata is required by or may optionally be placed in MXF files. System or structural metadata is about the structure of the file, e.g., the relationship of parts, whether the essence is stored as little or big endian, index tables that provide information on the essence (display size, compression algorithm, the time line of a media clip, etc.), size of a sector, where a new partition starts, etc.

Digital Object Type: Moving Image

Recommended Metadata Standard

MPEG/7, Version 10, 2004

Technical Metadata for Multimedia (MPEG-7), formally called the Multimedia Content Description Interface, is a multimedia content description standard, associated with the content itself. MPEG-7 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group). It is intended to allow fast and efficient searching. It does not deal with the actual encoding of moving pictures and audio (as MPEG-1, MPEG-2 and MPEG-4 do). It is intended to provide complementary functionality to the previous MPEG standards.

Link: http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm

Library of Congress[199] developed XML technical metadata schemas for their Audiovisual Prototype Project; these were widely implemented because of the lack of other schemas. Audio and video technical metadata schemas are under development by expert organizations.
Moving Image Collections (MIC) project [200].

Other Metadata Standards for Moving Image

MPEG/21 DIDL (2002) [201]: MPEG-21 aims at defining a normative open framework for multimedia delivery and consumption for use by all the players in the delivery and consumption chain. This open framework will provide content creators, producers, distributors and service providers with equal opportunities in the MPEG-21 enabled open market. This will also be to the benefit of the content consumer providing them access to a large variety of content in an interoperable manner.

MPEG-21 is based on two essential concepts: the definition of a fundamental unit of distribution and transaction (the Digital Item) and the concept of Users interacting with Digital Items. The Digital Items can be considered the “what” of the Multimedia Framework (e.g., a video collection, a music album) and the Users can be considered the “who” of the Multimedia Framework.

[198] http://www.digitalpreservation.gov/formats/fdd/fdd000013.shtml
[199] http://www.lov.gov
[200] http://mic.loc.gov/

[201] http://mpeg.chiariglione.org/standards/mpeg-21/mpeg-21.htm

The goal of MPEG-21 can thus be rephrased to: defining the technology needed to support Users to exchange, access, consume, trade and otherwise manipulate Digital Items in an efficient, transparent and interoperable way.

Digital Object Type: Document (Includes text documents, spreadsheets and presentations)

Recommended Metadata Standard

Document Metadata: document technical metadata for digital preservation (Chou & Goethals) Florida Digital Archive / Harvard University Library, 2009.

Extracting technical metadata from documents is essential as it can aid in characterizing the kinds of documents in our preservation collections; listing document properties that may hinder preservation (encryption, external fonts, etc); and providing requirements in selecting tools/facilities for document transformation including normalization and migration. In addition, document technical metadata can be used to verify the result of document transformations, ensuring the properties of the original document are preserved and properly transformed to the new document format.

"When it comes to document formats such as PDF, Word or OpenDocument Text, it has come to our attention that there is currently no technical metadata standard to follow." They "hope to develop a document metadata schema which is simpler and may be applied to document formats other than PDF. The document metadata schema may be expressed in XML or database form."

Link: http://fclaweb.fcla.edu/uploads/Lydia%20Motyka/FDA_documentation/documentMD.pdf

Draft implementation: http://www.fcla.edu/dls/md/docmd.xsd

Other Metadata Standards

No other widely adopted standards have been found in our study.

Digital Object Type: Executable

Recommended Metadata Standard

Preservation Metadata for Digital Collections, Section #5.6.01, National Library of Australia (1999)

5.6.01 Code Type and Version

Definition: The code type used to compile the executable and version.

Examples:
1. Compiled using Intel code executable for Windows 95 environment
2. Compiled using Perl script
3. Java version 1.2

Link: http://www.nla.gov.au/preserve/pmeta.html

Other Metadata Standards

No other widely adopted standards have been found in our study.

Summary Table of Metadata Standards

The table below indicates whether the recommended technical metadata standards meet our ten criteria.

	Criterion	TextMD	MIX	AES	MPEG/7	Document MD
1	Fit for purpose	YES	YES	YES	YES	YES
2	Open / non-proprietary	YES	YES	YES	YES	YES
3	Widely adopted	YES	YES	YES	YES	NO
4	Maintained / supported	YES	YES	YES	YES	YES
5	Documented	YES	YES	YES	YES	YES
6	Interoperable	YES	YES	YES	YES	YES
7	Format-specific and covers formats in scope	YES	YES	YES	YES	YES
8	XML-based	YES	YES	YES	YES	YES
9	Integrates with PREMIS	YES	YES	YES	YES	YES
10	Integrates with METS	YES	YES	YES	YES	YES

5.4.2 Provenance and Contextual Metadata

This section discusses the use of a preservation metadata standard to (a) describe aspects of digital objects in a blog, and (b) to manage and record repository actions that take place within an OAIS-compliant repository environment.

This will include provenance information regarding the source blog, pragmatic information surrounding the blog (such community characteristics as discussed in Chapter 4) and changes that take place within the repository in relation to metadata, object transformation, rights information, and/or policies.

Available Preservation Metadata Standards

Two available preservation metadata standards are:

1. PREMIS: http://www.loc.gov/standards/premis/
2. LMER: http://www.d-nb.de/eng/standards/lmer/lmer.htm

PREMIS

PREMIS is a Data Dictionary prepared by the PREMIS Working Group [202]. The Report provides a wealth of resources on preservation metadata. First and foremost is the Data Dictionary itself, a comprehensive, practical resource for implementing preservation metadata in digital archiving systems. The Data Dictionary defines preservation metadata that:

Supports the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context;

[202] http://www.loc.gov/standards/premis/index.html

Represents the information most preservation repositories need to know to preserve digital materials over the long-term;
Emphasizes "implementable metadata": rigorously defined, supported by guidelines for creation, management, and use, and oriented toward automated workflows; and
Embodies technical neutrality: no assumptions made about preservation technologies, strategies, metadata storage and management, etc.

The PREMIS Data Dictionary for Preservation Metadata version 2.1. is the current version and was published in January 2011.

The PREMIS system proposes implementation using a model of Agents, Events, Objects, and Intellectual Entities; put simply, **Objects** (e.g. digital objects, harvested blogs, storage files, individual formats) will undergo **Events** (e.g. migration, checksum, fixity check, virus check, validation) performed by **Agents** (e.g. software applications, repository managers). In PREMIS, this model describes the actions of a well-managed repository whose aim is long-term preservation of its resources. If implemented, PREMIS metadata will provide a detailed record of these preservation actions.

LMER

LMER is Long-Term Preservation Metadata for Electronic Resources, a schema supported by the German National Library [203]. For a workable strategy on long-term preservation of electronic documents, the compilation of suitable technical metadata is essential. Up to now, unfortunately no standard has been established for a corresponding metadata XML Schema especially in terms of long-term preservation. Thus, the German National Library with LMER provides an own Schema that is based upon a model of the National Library of New Zealand.

The German KOPAL project is using LMER along with METS profiles to deliver an OAIS-compliant solution called "The Universal Object Format" for archiving and exchange of digital objects. [204]

Meeting Selection Criteria

LMER

Table 5.2-6 Selection criteria for LMER

	Criterion	Met
1	Fit for purpose	YES
2	Open / non-proprietary	YES
3	Widely adopted	NO
4	Maintained / supported	YES
5	Documented	YES
6	Interoperable	YES
7	Format-specific and covers formats in scope	YES
8	XML-based	YES
9	Integrates with PREMIS	YES
10	Integrates with METS	YES

[203] http://www.d-nb.de/eng/standards/lmer/lmer.htm
[204] http://kopal.langzeitarchivierung.de/index_standards.php.en

PREMIS

Table 5.2-6 Selection criteria for PREMIS

	Criterion	Met
1	Fit for purpose	YES
2	Open / non-proprietary	YES
3	Widely adopted	YES
4	Maintained / supported	YES
5	Documented	YES
6	Interoperable	YES
7	Format-specific and covers formats in scope	YES
8	XML-based	YES
9	Integrates with PREMIS	N/A
10	Integrates with METS	YES

Both schemas meet our criteria but PREMIS is more widely adopted.

Examples of PREMIS in METS

Numerous examples of schemas demonstrating use of PREMIS in METS are available at http://www.loc.gov/standards/premis/premis-mets.html.

In the ECHO Dep Generic METS Profile for Preservation and Digital Repository Interoperability [205], special attention has been given to administrative and technical metadata, particularly on integrating the PREMIS data model and schema into METS (see Appendix D).

5.4.3 Rights metadata

Rights metadata are metadata documenting the rights holders, copyright status, permissions, agreements, terms and conditions, and licensing information associated with a resource. There are several **rights expression languages** (RELs) and other **metadata standards** that include fields for statements of digital rights. Below is an overview and a brief description of the main standards for rights management. These allow the expression of rights statements associated with a particular digital object or resource. Most are forms of descriptive metadata, which are aimed at imparting rights information to the users of a digital resource.

Rights Expression Languages

A Rights Expression Language or REL is a machine-processable language used for Digital Rights Management. Some of the most notable RELs are the following:

copyrightMD

CopyrightMD is an XML schema for recording characteristics that, taken together, help determine the **copyright status** of a resource. In 2004, California Digital Library (CDL) formed a short-term Rights Management Group (RMG) to advise on issues concerning rights protection and fair use (CDL, 2009). From 2005 through 2006, the RMG did an analysis of the functional requirements related to copyright metadata, identified key data elements for expressing copyright metadata, and

[205] http://www.loc.gov/standards/mets/profiles/00000015.html

formalized these elements into a prototype "proof of concept" schema, copyrightMD XML. copyrightMD is compatible with and can be used as an extension to the METS (Metadata Encoding and Transmission Standard).

METSRights

METSRights is an extension schema to the METS packaging metadata standard. It is an XML schema for documenting minimal administrative metadata about the intellectual rights associated with a digital object or its parts (Metsrights, 2011). METSRights is most often used to record statements to be viewed by professionals managing the content or to be displayed to end users viewing the content. It is not designed to be machine-actionable. It is divided into three principal sections, although the highest, root, level also has attributes which enables the specification of the kind of rights being described, e.g. copyrighted, licenced, public domain, contractual, or other. The three main sections are:

<RightsDeclaration> a broad declaration of the rights associated with a digital asset or part of a digital asset intended to inform the user community of these rights.

<RightsHolder> details of any person or organisation holding some rights to a given digital asset or part of a digital asset.

<Context> describes the specific circumstances associated with who has what permissions and constraints.

XrML

XrML is a proprietary method for securely specifying and managing rights and conditions associated with all kinds of resources including digital content as well as services. It underlies commercial Digital Rights Management applications (XrML, 2011). XrML has come to agreements with MPEG and other initiatives to enable them to use XrML as a basis for more specific rights language specifications, such as MPEG21-Part 5: Rights Expression Language (MPEG21, 2011).

Open Digital Rights Language (ODRL)

Open Digital Rights Language (ODRL, 2011) Initiative is an international effort aimed at developing and promoting an open standard for defining a model and vocabulary for the expression of terms and conditions over assets. ODRL provides flexible and interoperable mechanisms to support transparent and innovative use of digital content in publishing, distribution and consumption of digital media across all sectors and communities. It is used in Digital Rights Management and open content management systems. It also provides the semantics to express policies which might be enforced by a machine-actionable DRM system. ODRL is a rights metadata scheme that covers transaction activities. Transaction metadata is for materials being sold or licensed today.

Creative Commons Rights Expression Language (ccREL)

Creative Commons Rights Expression Language (ccREL) is a proposed Rights Expression Language (REL) for descriptive metadata to be appended to media that is licensed under any of the Creative Commons licenses. According to the draft submitted to the W3C, it is to come in the forms of RDFa for (x)HTML pages and XMP for standalone media.

Creative Commons provides a range of standardized **digital licenses** that can be associated with or embedded in open access web resources (Creative Commons, 2011). It is a form of licensing which

enables copyright holders to grant some of their rights to the public while retaining others through a variety of licences. The licences were developed in recognition of the fact that many rights holders do not wish to restrict the use of their materials as rigidly as the default copyright protections and may in fact wish to encourage re-use of their creations. Creative Commons allows creators to generate licences for their materials very simply, by completing an online form. The Creative Common Licenses includes three major characteristics:

- ✓ Permissions - rights granted by the licence.
- ✓ Prohibitions - things prohibited by the licence.
- ✓ Requirements - restrictions imposed by the licence.

Encoded Archival Description (EAD)

A widely adopted standard for encoding archival finding aids modelled upon the International Standard Archival Description (General) (EAD, 2011).

EAD includes two elements relevant to IPR:
<accessrestrict> and <userestrict>

Metadata Initiatives for Rights Management

IPRs can also be managed by metadata standards and initiatives such as:

Dublin Core

Simple Dublin Core has 15 elements which may be used to describe a resource (DMCI, 2011). One of these is specifically for the description of IPR rights attached to one or more digital objects: <dc:rights>. This field can be use 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.

Qualified Dublin Core

Qualified Dublin Core extends the 15 core descriptive elements, providing a more granular metadata structure.

Elements relevant to rights are:

<dcterms:accessRights> Information about who can access the resource or an indication of its security status.

<dcterms:dateCopyrighted> Date of a statement of copyright.

<dcterms:license> References a legal document giving official permission to do something with the resource, preferably via a URI. However, this might also be a hard-copy deposit or donation agreement.

<dcterms:rightsHolder> A person or organisation owning or managing rights over the resource.

Preservation Metadata Implementation Strategies (PREMIS)

The PREMIS Data Dictionary (PREMIS, 2011) includes semantic units for Objects, Events, Agents and Rights. It provides these elements along with information about how to apply these in order to support the long-term preservation of digital objects. The **Rights entity** takes the form of a structured permission statement linked to a digital object, presumably the object being preserved.

Type of Rights we need to manage in BlogForever

The types of rights, and mechanisms of managing these rights within BlogForever will be further investigated in the BlogForever deliverable D3.3 Digital Rights management Policy. Here we mention topics that we will be further investigating in this context to determine what types of rights a repository might need to manage within the web archiving and weblog archiving context.

We would need to reflect the thoughts from the general discourse on Digital Rights Management as regards Web Preservation. Thoughts from the Internet Memory Foundation 2010 survey on Web archiving might also provide insight into this topic.

In particular, the aims of digital rights management should be clearly specified: for example, rights management policies should be conscious of protecting the public access right to information, that is, to facilitate access to information and the freedom to express views. On the other hand, rights management policies must be designed to protect the rights of content creators: for example, ownership and intellectual/digital property rights, privacy and surveillance rights, human rights and other civil rights (e.g. right to delete and forget). In conjunction, policies should be aware of issues surrounding the protection of content managers (with respect to liability, accountability, justification).

The concerns as regards weblogs are manifold as the blogs often contain third party material and often involves multiple authors and other content providers (e.g. the technological setup of the blog might be handled by another party). The rights (e.g. licenses and patents) associated to selected object types and format will also need to be considered in developing such policies. These topics must also address potential conflicts that might arise between question of rights and question of preservation: for example, how do we resolve the problems that exist between the right to delete information and the responsibility to protect historically, culturally, or politically relevant information? The mechanisms that must be put in place to monitor changing rights and policies as well as verifying that policies are upheld will also require special attention.

In BlogForever, we will address this in detail by conducting a survey of legislation and policy and through case studies of existing web archiving projects and weblog content creators and managers to define the problems and challenges/opportunities.

Selecting the rights metadata for BlogForever

As reported in the previous section, many efforts are concerned with rights expression languages and metadata related to intellectual property rights and permissions. However, only a small body of work addresses rights and permissions specifically related to digital preservation.

PREMIS 2.0 includes a completely revised and expanded Rights entity. The Rights entity in PREMIS 2.0 and later in version 2.1 is intended to support an automated process that determines if a particular preservation-related action is permissible in regard to an Object or set of Objects within the repository, as well as to record important information about the permission. PREMIS 2.0 can be used to express three forms of intellectual property rights: those established by **copyright**, those established by **license**, and those established by **statute**. The Rights entity defines metadata applicable to all three forms of rights statement, such as identifiers, the nature, scope, and characteristics of the rights granted to the repository, the Object(s) to which the rights apply, and the Agents responsible for granting or administrating the rights. In addition, the new Rights entity

defines metadata specific to copyright-, license-, and statute-based intellectual property rights. The result is a deeper, more nuanced description of rights in a digital preservation context.

Right metadata through PREMIS in METS

As the push for long-term access to digital information increases, a growing number of organizations are using PREMIS in METS (Metadata Encoding and Transmission Standard) to record provenance and other information that supports sustained access. The METS schema (METS, 2011) is widely used by digital repositories as a packaging mechanism for objects and their associated metadata. A number of questions have emerged as to how the PREMIS Data Dictionary and schema should be used in conjunction with METS. The Maintenance Activity has convened a group of experts to develop a set of guidelines and recommendations for using PREMIS and METS. A working draft of their findings is now available online (PREMIS and METS, 2008).

METS schema specifies administrative metadata section (amdSec) with the following elements:

1.1 techMD
2.1 rightsMD
3.1 sourceMD
4.1 digiProvMD

PREMIS Rights metadata should be used in the “rightsMD” METS section. If using all PREMIS units together the entire package goes in digiProvMD with the <premis> element as a container. An example on how PREMIS can be used in METS to express rights metadata is shown in Appendix E. Further discussions of rights metadata will take place in the BlogForever deliverable D3.3 Digital Rights Management Policy.

5.5 METS: a Wrapper for Recommended Metadata

In the BlogForever project we have decided to use METS as the standard to keep all the metadata needed for the blogs archive. In this document we will describe a draft idea of how to use METS together with other formats identified in the previous sections.

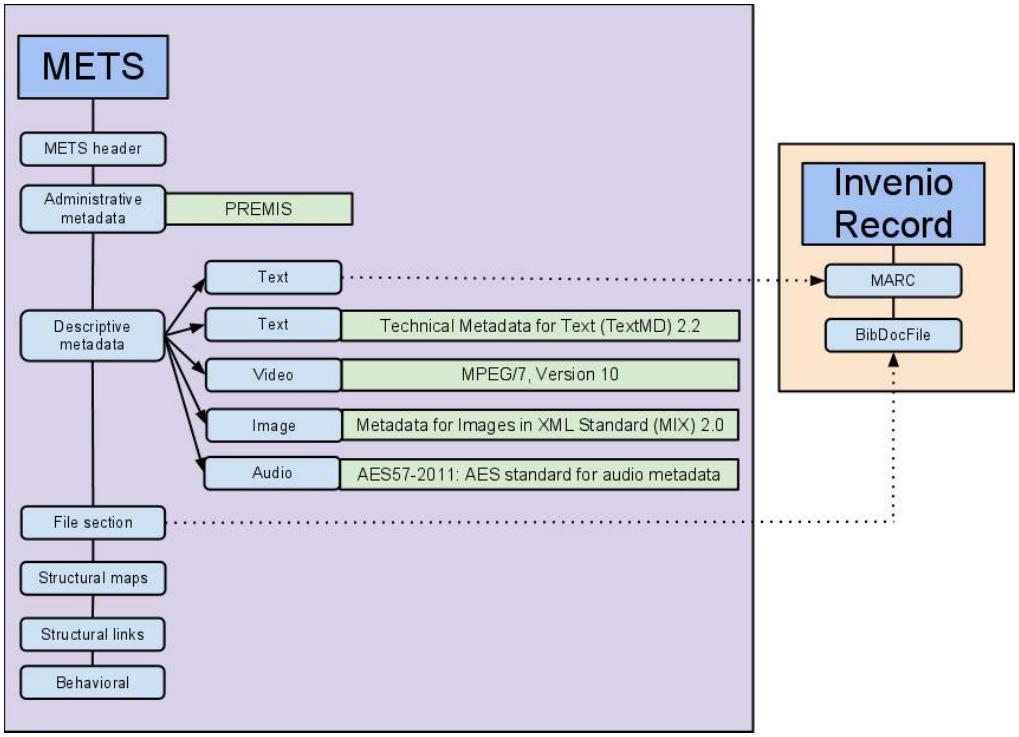


Figure 5.5-1 Proposed METS structure.

An Invenio record has two main components: MARC metadata and the associated files. The associated files are kept using the BibDocFile module. We would like to keep Invenio using MARC for records and also store a METS object (as a xml file) with the record, using BibDocFile. This METS file has several sections. We will now focus on the administrative metadata section (*amdSec*), the descriptive metadata section (*dmdSec*) and the file section (*fileSec*).

METS supports multiple descriptive metadata sections. In task 3.1.a, UL proposed a metadata standard for every type of content. Since METS supports other metadata standards to be embedded in *dmdSec* section, the metadata will be included in METS embedding the xml code of the correspondent standard. In addition to embedding, METS also supports linking[206]. We will use this technique to link the METS object to Invenio's MARC. There, a clean plain text version of the text content will be stored, as well as a copy of the metadata fields that we consider interesting for indexing, but keeping the full version of the metadata in the METS file.

In the *fileSec* there will be references to the files stored next to the METS file, also in BibDocFile, and the administrative metadata will be stored in the *amdSec* section using PREMIS. The generic schema of this architecture can be found in Figure 5.5-1 and a specific example of a post consisting in text and two pictures can be found in Figure 5.5-2.

Our proposal is to have 4 different kinds of Invenio records (Blog, Post, Comment and Page), treated as "equal citizens" in the system, but, of course, differentiable during search, with different possible ways of being displayed and displayed differently in the user interface. The reason for having these as independent records is that, in the data model resulting from D2.1, the four entities all have "content" (text and multimedia), but the rest of the metadata is different. The fact that these are at the same level does not mean that the hierarchy information is lost. The records can be linked using MARC tags. This is more flexible than a vertical hierarchy structure that could result in a huge METS object.

[206] An example of how to embed and link external metadata into METS this can be found in CERN's collaboration in 3.1.a.

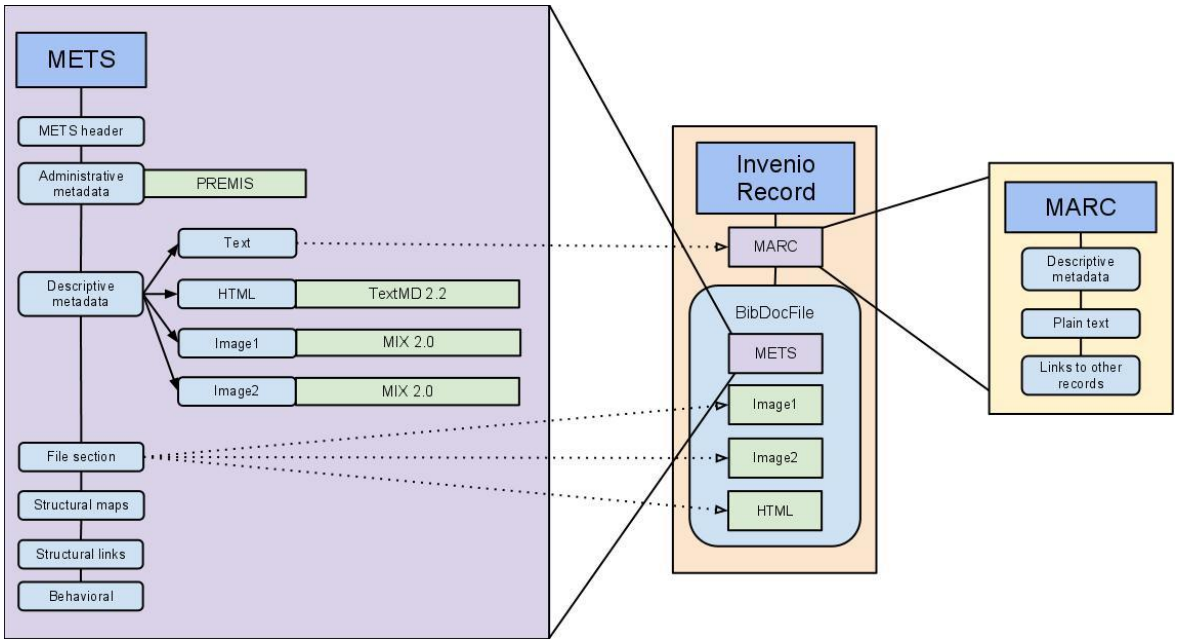


Figure 5.5-2 Example of a Post as a record within Invenio.

6 Repository Audit Standards

In this chapter, we investigate the BlogForever repository's ability to meet repository audit standards. As there is no repository yet in existence, this examination can only be anticipatory in nature. There has been a long history of initiatives to provide guidance for the certification of repositories as trusted digital repertories. To name a few:

- Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) (Ambacher et al. 2007)
- Catalogue of Criteria for Trusted Digital Repositories (nestor Catalogue) (Dobratz et al. 2006, 2009)
- DCC and DPE Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) (Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), 2007)
- DINI-Certificate Document and Publication Services (DINI AG Elektronisches Publizieren, 2006)
- Data Seal of Approval (Sesink et al. 2008)

All of these were developed cognizant of the international standard reference model for an Open Archival Information System[207] (ISO 14721:2003). The TRAC checklist and the nestor catalogue were subsequently approved as international standard ISO 16363 and German National Bureau of Standards DIN 31644. Most of these approaches comprise a "tick-the-box" checklist methodology for measuring repository trustworthiness, except DRAMBORA, which extends the check-list methodology to a risk management approach of identifying risks and estimating their impact as a means of measuring repository trustworthiness.

Here we discuss the BlogForever repository as a trustworthy repository, first on the conceptual level of the OAIS (Section 5.1) and, then, briefly with respect to DRAMBORA (Section 5.2).

6.1 The BlogForever Repository and the OAIS

The purpose of this section is to introduce a workflow developed at the University of London to address the relationship between the reference model for the Open Archival Information System (OAIS) and the BlogForever repository. Early on in the project it was already observed that the model may not be adequate as a foundation for a repository of web archives (Kim and Ross 2011a). For example, in referring to the OAIS mandates:

1. There is seldom explicit negotiation for the acquisition of web pages. While there are options to try to prevent a spider from harvesting a selected page, this is not consistently applied nor is it clear that this is an ethical policy. Sometimes the wishes of the website owner are overridden on the basis of legal mandates issued on a national level: this is only reasonable if it is assumed that 1) web pages are assumed to be instances of "printed publication", 2) it is possible to determine the legal jurisdiction of a "published" web page.
2. The lack of negotiation means that the archive's right to manipulate the webpage for preservation purposes becomes questionable, and the archive holding's integrity can be put at risk (the creator can request material, even parts of a page, to be deleted at any time).
3. Even when permission is negotiated, the page itself is often acquired using web spidering and archiving technology (e.g. Heritrix, Archive-It): that is, the relationship between the archived copy and the copy at the time of creation can be unpredictable.
4. There is rarely any information package submitted by the information producer, that is, the consistency of adequate information provided across items may vary greatly, especially in an environment like the web where many different technologies are utilised.

[207] http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

5. There is no explicit notion of a designated community articulated by the web archives.
6. The boundaries of what constitutes an “intellectual entity” (e.g. in the sense that Shakespeare’s Hamlet is an intellectual entity) is unclear in the weblog environment (cf. discussion in Chapter 1 on “identity”).

The task began life as an examination of OAIS compliance[208]. Following discussions with UG in February 2012, a question was raised about the need for BlogForever's "full compliance" with the model, particularly in the light of the expense of conducting an OAIS assessment (the results of which can sometimes be disappointing). Following further discussion at the 3rd Consortium Meeting in Berlin it was agreed that: "According to the DoW, BlogForever will utilize the Open Archival Information System (OAIS) reference model as a conceptual guidance for its weblog digital repository construction and management."

The discussion here comprises:

- A set of *workflow instructions* which, if followed, would enable the repository to preserve digital objects in an OAIS-like manner.
- A draft set of *preservation service requirements* which, if developed further and implemented, would equip the repository to preserve digital objects.
- Observations on other OAIS functions.
- A description of the three OAIS *Information Packages* and what we anticipate they will look like in BlogForever.
- A reiteration of the BlogForever stakeholders identified in D4.1 and how they map to OAIS Actors.
- A *workflow overview*, presented as a chart that maps to the high-level OAIS model.

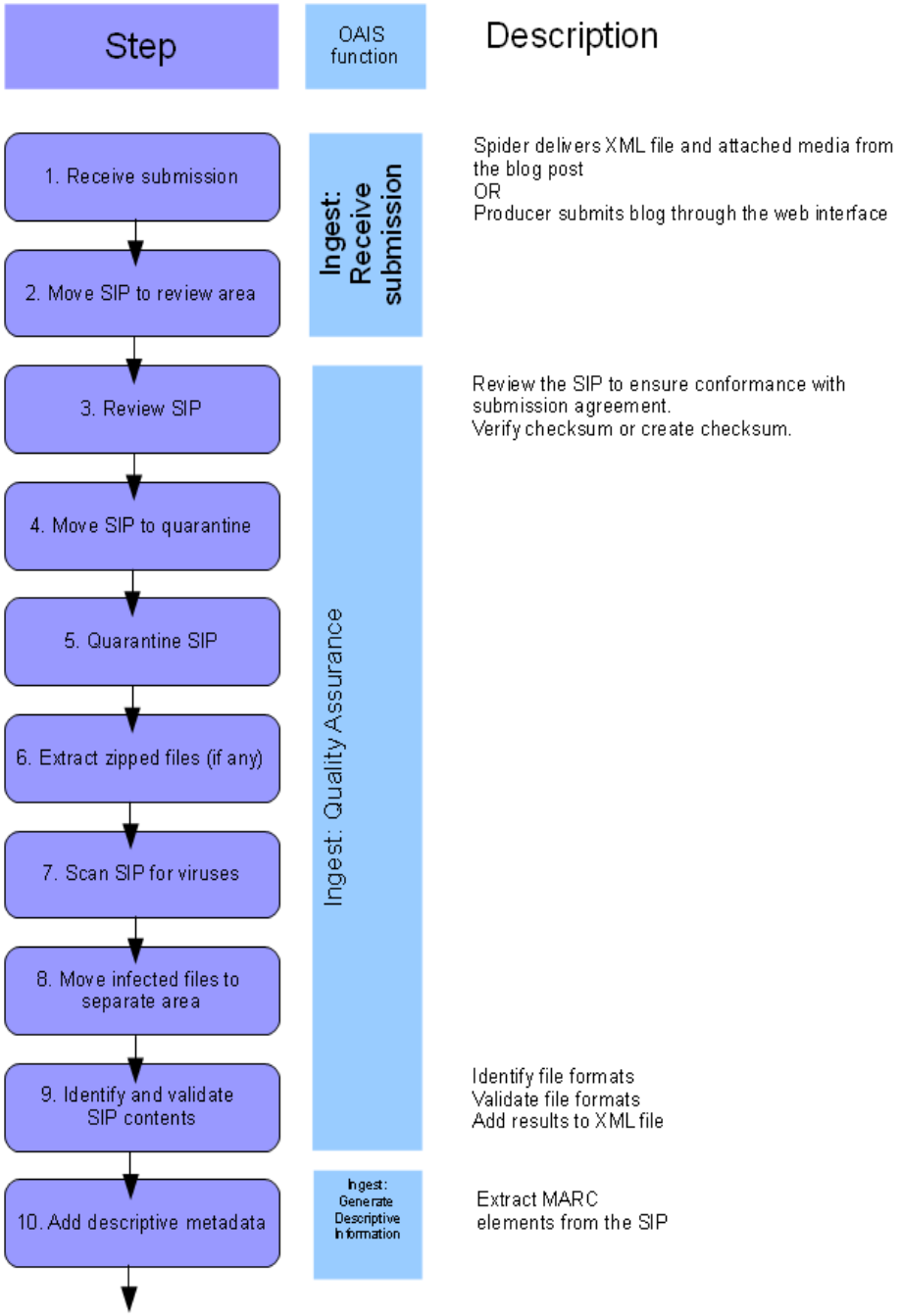
The OAIS model is used as a conceptual framework to build a workflow selectively to be applicable to the weblog repository context. The focus is on establishing a practical set of functions and requirements that helps the BlogForever software and repository to perform preservation, rather than a strict examination of "compliance" with OAIS.

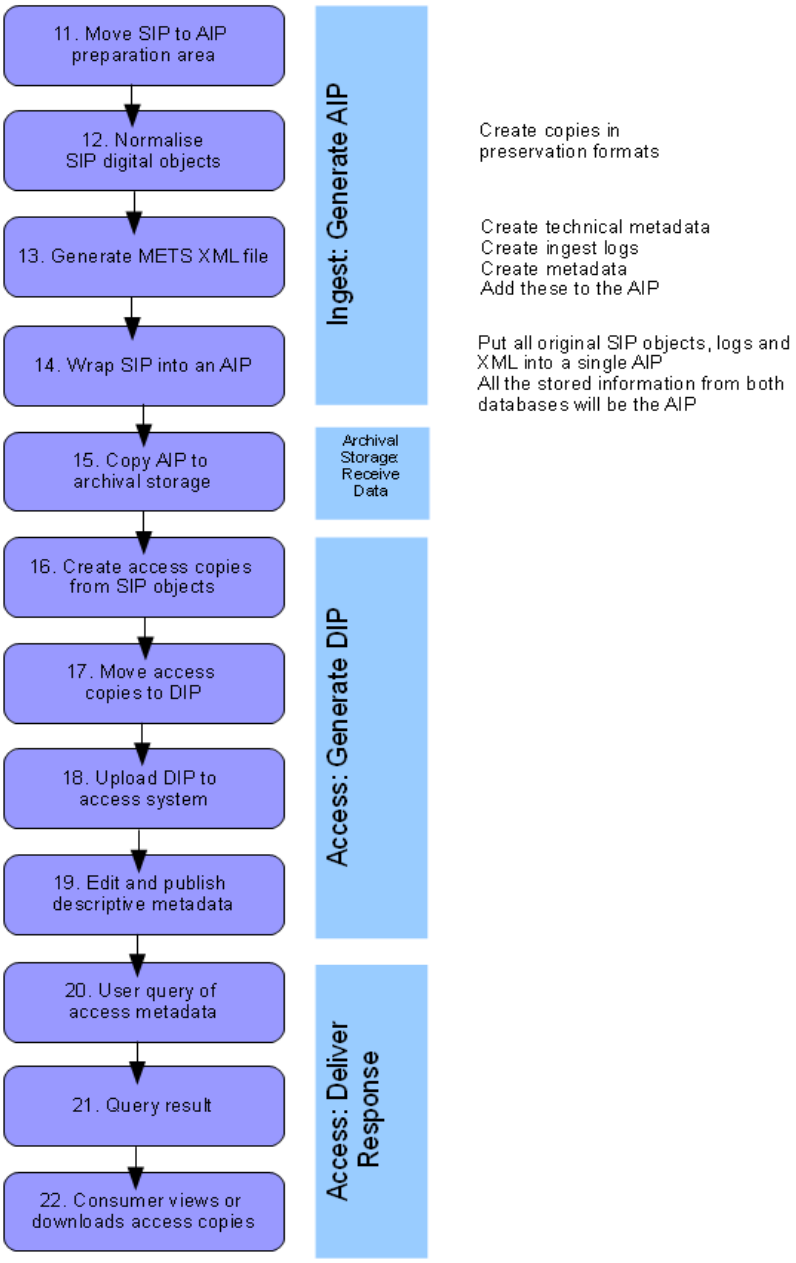
6.1.1 Proposed Workflow

We begin by introducing the general proposed workflow of BlogForever.

[208] The original wording for the task was "Examination of OAIS compliance: A document describing how the current data model, and repository components map to the OAIS model."

BlogForever
Proposed Workflow
Version1.0
UoL, March 2012





6.1.2 Preservation Service Recommendations

This section comprises some preservation service recommendations that should be taken into account during the design of the BlogForever repository to perform four of the six functions defined by OAIS. Some of the functions may already be met or partially met by Invenio's existing capability. It should be stressed this section represents the first stage in drafting such features.

This report uses the OAIS framework *selectively*. We concentrate on the four core functions that we would expect to find in a repository workflow to enable preservation to take place. These four functions are Ingest, Data Management, Archival Storage and Access. The two remaining functions, Preservation Planning and Administration, are discussed in Section 6 below.

It may be interesting to note that the majority of recommendations are related to the Ingest stage. This matches the perception of Adrian Brown (Parliamentary Archives) who has said that "Ingest accounts for up to 90% of digital repository activity". [209]

Some of the recommendations are already being met or partially met by the requirements outlined in D4.1 User Requirements deliverable, especially with regard to the Access function. Where other D4.1 requirements appear to be relevant to each recommendation in this exercise, we note them under the "see also" reference.

Ingest Recommendation

A large part of the ingest function would be performed by Invenio's WebSubmit module along with other mechanisms. It would offer:

- An interface to the Crawler (spider), usable by the repository managers
- A submission interface, allowing Producers to submit blogs in the forms of SIPs

1: Receive Submission

Workflow steps 1-2.

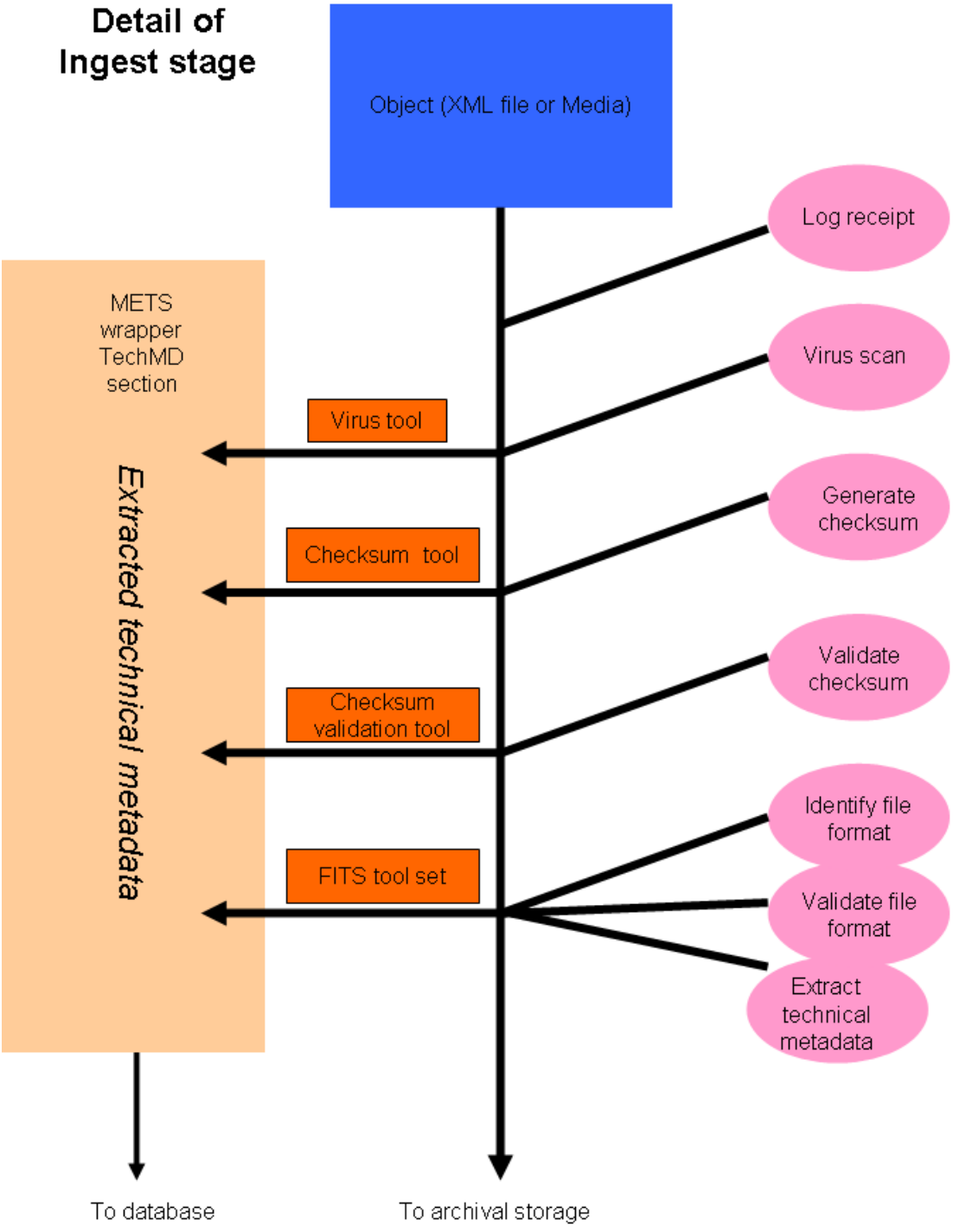
The repository should provide a method to submit blogs into BlogForever repository.

The submission should be done through one only entry point. Content managers (or normal users, if the administrators decide so) would push new blogs in which is the Invenio WebSubmit module.

In BlogForever, the Submission Information Package will probably be a METS wrapper which contains the original XML data as crawled by the spider, along with MARC and MIX metadata, and links to locally-attached files.

See also
FR15 - Selection of blogs to archive
FR37 - Web portal
IR2 - Capturing is possible for various platforms
PR1 - Amount of blog posts to capture
IR8 - Digital Object Identifier
FR48 - Crawler/Spider Support Platform Flexibility
FR49 - Support Different Versions of Blogging Software
FR54 - What to Archive: Text and Comments
UI34 - Simple Submission by Authors
UI35 - Workflow to Manage Blog Submissions

[209] "Making sense of digital collections – ingest, characterisation and workflows in archives". DPC event *Digital Preservation: What I Wish I Knew Before I Started*, January 2012.



2: Quality Assurance

Workflow steps 3-9.

The repository should perform validation of the transmitted content to ensure that the transmission was successful and that the content is eligible for admission to the repository. This validation may include:

1. Ensuring BlogForever can support the file formats in the SIP
2. Verifying or creating a checksum
3. Quarantining the SIP
4. Extracting any compressed or zipped files
5. Scanning the SIP for viruses, and taking appropriate action
6. Identifying file formats
7. Validating file formats

We anticipate that checksums, virus scanners and automated file validation tools will be used, and that any useful outputs from the above actions will be added to the ingested METS file.

See also
FR47 - Data integrity

3: Generate descriptive information

Workflow step 10.

The repository should create discovery metadata for search and retrieval of the blog.

Functions to create and edit the descriptive metadata are already implemented in Invenio. If the producer supplies metadata with their SIP, it's acceptable for the repository staff to enhance this metadata and create an "Updated SIP" in OAIS terms.

We anticipate the descriptive information will be held in a MARC schema and be added to the ingested METS file.

See also
DR17 - Metadata for blogs

4: Generate AIP

Workflow steps 11-14.

The repository should transform a submitted SIP into an archival AIP. The transformation method may vary according to the needs of the blog and the formats and media within it.

We anticipate the creation of an AIP will involve storing the content in two different databases.
A copy of the MARC metadata (i.e. the descriptive information) would be stored in the "Main Storage Database" where it would be processed in order to extract information and retained for further processing and output,

The original METS file *as submitted* would be stored in a separate "Ingestion Database" for preservation purposes. This version of the AIP will later be rendered as a DIP.

The main "transformation" consists of storing the data in multiple places, with the ability to pull it together again through the use of a UID. There would be a submission ID stored in the METS header, so all the stored data in both databases would be characterized by that UID and would be the OAIS AIP.

There may also be a recommendation for normalisation or format-shifting (i.e. migration) of the media attachments found in blogs, such as text and images. This depends on:

- Whether the media is captured at all (the spider may not always harvest it)

- What formats we anticipate finding in blogs. See Section 2.1, *Scope of Formats for the BlogForever Repository*.

See also
RA2 - Correct information in the archive
IR8 -Digital Object Identifier
FR51 - UTF-8 - The Default Character Encoding

5: Co-ordinate updates

The repository must should move AIPS into archival storage, and store descriptive information in the database.

This OAIS recommendation is simply describing the automated actions of a repository. Invenio already has tools (embedded in the various modules) for these tasks.

Archival Storage Recommendations

6: Receive data

Workflow step 15.

The repository should move an AIP into permanent storage.

See recommendation 4: Generate AIP above; this is performed by the functions of the two databases, the Main Storage Database and the Ingestion Database.

See also
PR2 - Storage data concurrently

7: Manage storage hierarchy

The repository should implement a backup strategy. Suitable hardware and procedures should be needed to ensure the appropriate level of protection for the AIP. This strategy may include the following elements:

- Backups stored on a server (e.g. a RAID server).
- Provision of multiple redundancy in case of hard disk failure.
- Tape backups are stored offsite.
- Provision of error checking and error logs for media failure.
- Provision of operational and usage statistics.

This strategy is not explicitly included in this report's proposed workflow.

It's worth stating that these storage recommendations are service-related recommendations and not software-related recommendations. BlogForever is producing software and not a service. This OAIS requirement might be better expressed as a recommendation to the final administrators of the system better than a recommendation for the design.

See also
CS1 - Amount of archived blogs
CS2 - Amount of blog posts per day

8: Replace Media and Migration Strategy

The repository should be capable of reproducing the AIPs over time. This includes error checking for media failure in storage, but also the migration of file formats when necessary.

The migration strategy is a process which validates data and migrates it when necessary. It may include the following stages and elements:

- Assess risks to file formats.
- Preserve at-risk formats by taking correct actions (see preservation strategy).
- Evaluate outcomes of migration, run error-checking procedure.
- Update preservation metadata.

This strategy is not explicitly included in this report's proposed workflow.

See also
SP2, Mechanisms to avoid data loss
SM1 - Migration/Updating without down time
SM3 - Data export for migration
OP1 - Versioning

9: Error-checking

The repository should provide assurance that the storage and data transfer process has not corrupted the AIP. This action may include the following elements:

1. Run error-checking procedure.
2. Carry out periodic obsolescence checks.
3. Validation of blog data.

This strategy is not explicitly included in this report's proposed workflow.

See also
SM4 - Compliance with nagios and cacti monitoring software

10: Disaster recovery

The repository must duplicate the contents of the archive and store the copies in a remote facility.

This strategy is not explicitly included in this report's proposed workflow.

See also
RA1 - Recovery of the system

Data Management Recommendations

This function is about the maintenance of the BlogForever repository database and its administrative functions. It includes updating the descriptive metadata catalogue.

None of these requirements are explicitly included in this report's proposed workflow.

11: Administer database

The repository will have a database which contains descriptive information and system information. The function must maintain its integrity.

12: Perform queries

The repository database should perform queries that can locate and retrieve blogs in response to requests.

13: Generate report

The repository system should create reports (e.g. on size of holdings in the archive, or usage statistics).

See also
FR3 -Descriptive statistics for the archive
FR5 - Descriptive statistics for a single blog or blog post
FR18 - Analyze the network structure of blogs

14: Receive database updates

The repository system should add, modify or delete database information in response to updates, such as ingest or access requests.

See also
FR23 - Detection of duplicates
OP3 - APIs for developers

Access Recommendations

15: co-ordinate access activities

See Workflow step 20.

The repository should provide a user interface to the archive holdings. Invenio has already its own web interface.

See also
EI1 - API for external clients to query data
EI2 - Data access/export as XML
EI4 - Accessible via search machines
EI5 - Export as CSV
UI1 - Web Interface

16: Generate DIP

Workflow steps 16-19.

The repository should allow an AIP to be converted into a DIP automatically. This involves copying an AIP from archival storage, adding descriptive information as needed, and updating the database.

The repository team have not designed this specific stage at time of writing. However, we consider that to produce a DIP it should be possible to retrieve the original METS file from the Ingestion Database, enrich it with extracted information, and export it to the designated community.

See also
FR4 - Blog export
FR17 - Print/Export as PDF, JPEG, etc.

17: Deliver Response

Workflow steps 20-22.

The repository should deliver responses to consumers.

The search web interface, search engine user interface, and community tools are already implemented by Invenio. They should need to be extended to satisfy the project's specific requirements.

The following user requirements, already defined in D4.1, are directly relevant to Access requirements as understood by OAIS. In one sense all the requirements in D4.1 are valid for access.

Search and retrieval functions:
FR8 - Topics (Categories) for blogs and blog posts
FR13 -Keyword / metadata search
FR14 -Full-text search
FR16-Search by author
FR26 - Context-sensitive search by keyword
FR30 - Extract bibliographic metadata from blog contents
FR31 - Define important blogs and filter junk
FR34 - Topic/Subject detection
FR35 - Detection and ranking of the originality
FR36 -Memetracking and trend detection
FR37 -Web portal
FR38 -Multidimensional indexing
FR41 - Retrieving semi-structured information
FR43 -Access to content in a harmonized way
FR44 -Advanced searching
FR45 -Personalized filtering services

Data requirements:
DR2 -URI and metadata for referencing / citing
DR4 -Author of the blog, blog post, comment
DR5 - Tags of the blog or blog post
DR6 -Metadata for captured contents
DR7 -Date / timestamp for creation and capturing
DR9 -Connections / links
DR11 -Differentiate between blog and blog post
DR13 -Comments
DR16 -Search keywords
DR17 -Metadata for blogs

IR3 - Export data using OAI-PMH protocol and Dublin Core schema
IR4 - Expose parts of the archive via OAI-PMH based on specified criteria
IR5 - Connection with federated search engine dbwiz

User interface requirements:
UI1 - Web interface
UI5 - Citation is presented prominently
UI6 - Latest posts
UI7 - Tags for blogs and blog posts
UI8 - Overview with metadata and summary
UI9 - Network view for topics, blogs, posts, authors, etc.
UI11 - Historical / chronological view on a blog
UI15 - Search interface
UI16 - Easy to learn / intuitive

UI21 - Archive content is clearly stated as such
UI23 - Categories/Topics are shown in different tabs
UI27 - Dynamic network view on topics, blogs, posts, etc.
UI30 - Creation of a Community of Providers and Recipients within the Archive Platform

6.1.3 Other OAIS Functions

The main focus of this report is suggesting basic preservation actions and a preservation workflow for the BlogForever software and repository. The aim of the document is to show how the BlogForever software and platform could easily incorporate a workflow that indicates OAIS compliance within four important OAIS functions.

Requirements for the two remaining OAIS functions, Preservation Planning and Administration, are presented here in a draft form. Some possible requirements within these functions are suggested, but these requirements are not as yet incorporated within our draft workflow.

Preservation Planning Function

In the course of our analysis and comparison with the OAIS framework, we formed the impression that this OAIS function is not clearly identifiable as a repository function which Invenio could own, and the requirements are high-level policy and management functions.

Some of the requirements within this function are preservation policy areas that would presumably be governed by the preservation strategy, which will be delivered by Workpackage 3. Since this function is probably not something Invenio will perform, we are (for the time being) designating this as the "The BlogForever Preservation Service".

However, it is not clear which entity would own this function after *completion* of the project. In OAIS, this function is usually owned by the Management entity: "Management is the role played by those who set overall OAIS policy as one component in a broader policy domain." The Deliverable D4.1 on User Requirements, while it has clearly identified stakeholders that map clearly to OAIS entities Producers and Consumers, has not yet identified a similar stakeholder who might act as Management. If BlogForever needs to identify a Management entity, it is possible this will be defined as part of the activities arising from WP6. For further observations on the ambiguity of Management, see section 8.

Monitor Designated Community

Description of the requirement: The BlogForever Preservation Service must monitor the user community. It will interact with Consumers and Producers of the blogs to identify any changes in what they require from the service, and remain aware of available product technologies that would help meet these requirements. Monitoring could take place via surveys, workshops, or a review process.

Develop Preservation Strategies and Standards

Description of the requirement: The BlogForever Preservation Service must develop and review preservation strategies. The strategy will devise and implement a method of identifying at-risk content; it will build a knowledge base of information required to support digital objects; it will have an understanding of the significant properties of file formats; and propose a method of preserving the content (such as migration).

Develop Packaging Designs and Migration Plans

Description of the requirement: The BlogForever Preservation Service will implement preservation strategies in stages likely to involve stages such as the following:

- Migration of blog content to a format that will preserve it
- Creation of preservation metadata to document actions performed on the blog's digital objects. (This will probably be done using the PREMIS standard.)
- Create fixity information
- Create a written agreement (between BlogForever and its Producers) that describes the terms of service
- Create written procedures for how to build or enhance metadata (probably using METS and MARC)
- Create written procedures for preservation of blogs transferred to archival storage.

Monitor Technology

Description of the requirement: The BlogForever Preservation Service will follow digital technology to identify any factors which may cause obsolescence within the archive and prevent access to the archive of blogs. This technology-watch function needs to keep abreast of emerging technologies.

Administration Function

We anticipate this function will come to be embedded in the entire BlogForever system. Most of the Administration requirements described below will probably come to be owned by Invenio as the service develops. Invenio may need to undertake more development to offer all the OAIS-requested tools and their corresponding web interfaces. There is also a dependency on the results of other work packages in the BlogForever project. For the time being, these requirements are likewise presented in a draft form, with a view to including them in a later iteration of the workflow.

Negotiate Submission Agreement

Description of the requirement: 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 current thinking on rights management is that there is some scope for adopting a mechanism similar to the Creative Commons automated licence.

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 case 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.

See also:

LR1 Copyright laws

LR2 Privacy laws
LR3 Additional national laws
LR4 License of the content
LR5 Open source software license is preferable
DR1 Rights and licenses
DR3 Disclaimer
FR6 Processing of licenses

Manage System Configuration

Description of the requirement: The repository system will maintain its integrity through a series of audits, statistical logs, change requests and reports.

Customer service

Description of the requirement: The repository will manage customer accounts and bill them automatically.

See also:

FR25 – Paid access/Billing system
FR40 Billing system

Archival information Update

Description of the requirement: The repository will operate an administrative function that allows parts of the system to update other parts.

Audit Submission

Description of the requirement: The repository will deliver a means of automatically validating a SIP.

Activate Requests

Description of the requirement: The repository will keep a record of event-driven requests.

6.1.4 Information Packages

OAIS identifies three types of Information Package, SIP, AIP and DIP. This section describes how we think these packages will look and behave in the BlogForever repository.

Submission Information Package

For BlogForever, the SIP will be a harvested blog created by one of the many content producers.

The SIP can arrive in many ways:

1. Delivered to the BlogForever Repository by the crawler / spider
2. Submitted by the blog owner or author
3. Submitted by another producer, for example a repository of digital content or another blog archive

SIP Contents

This virtual container will be the SIP. At time of writing, the current thinking is that it will be rendered in METS, containing descriptive and other metadata, and links to the digital objects.

Table 6.1-1: comparison of BlogForever and OAIS terms for SIP

In BlogForever	OAIS term
Metadata from the blog. See the data model. This would include all the constituent parts in the data model except possibly the "Categorised Content". It will be parsed XML data submitted from the spider crawl or provided directly by the data owner and will be stored in a repository database.	*Content data objects*
Digital objects - all the Categorised Content of a blog (see data model), or "media". CERN and others are currently thinking of these as "attachments" to the blog. Multiple file formats are possible - image and text formats will be common.	*Content data objects*
Discovery metadata. Probably will be in MARC and extracted from the blog by CERN.	*Representation Information*
Additional provenance and Context: Crawl logs from the crawler, or information supplied by the producer. [Awaiting feedback from Cyberwatcher; it remains to be seen if the crawl logs contain any provenance metadata of value (most harvesting engines discard them).]	*Context Information and Provenance Information*

Archival Information Package

This will be a derivative of the SIP that has been manipulated by BlogForever to make it suitable for preservation. The AIP is stored in the repository. Technically speaking, it is kept in multiple places.

AIP Contents

Table 6.1-2: comparison of BlogForever and OAIS terms for AIP

In BlogForever	OAIS term
Metadata from the blog. See the data model. This would include all the constituent parts in the data model except possibly the "Categorised Content". This will be the XML data processed by CERN.	*Content data objects*
Digital objects - all the Categorised Content of a blog (see data model), or "media". For the AIP, these media files may need normalisation or format-shifting.	*Content data objects*
Technical metadata. This information will describe the technical aspects of the deposited and archival versions of the blogs. It is anticipated BlogForever will use the METS metadata schema for this, including TextMD for text, MIX for images, AES for audio, MPEG/7 for moving images (see task 3.1.A for more detail).	*Representation Information*
Discovery Metadata to locate and retrieve the blog. This will be created by BlogForever administrators at point of ingest. This will be done using MARC XML.	*Representation Information*
Rights metadata to describe the rights associated with the blog. This will be created by BlogForever administrators. [Rights Schema not yet defined - see Work Package D3.3].	*Representation Information*

In BlogForever	OAIS term
Provenance metadata to describe the content history, including its origins. This might include harvesting logs from the spider and any other useful information about the crawl. It may also include information supplied by producers.	*Preservation Description Information*
Preservation metadata, including any metadata about future migrations and transformations of file formats or other content. This will be created and maintained by BlogForever administrators at point of ingest. The PREMIS standard will be used.	*Preservation Description Information*
Fixity information - to authenticate the digital objects. This will probably consist of running a checksum program.	*Preservation Description Information*
The AIP might be a blog inside a wrapper format, so there would be a requirement for some metadata about the wrapper too. The packaging information could also be a METS wrapper which encloses the entire blog post content and its metadata.	*Packaging Information*

Dissemination Information Package

This is a version of the blog that is intended for use by BlogForever consumers. The DIP version will be suitable for access by the web interface.

DIP Contents

Table 6.1-3: comparison of BlogForever and OAIS terms for DIP

In BlogForever	OAIS term
Digital objects - A rendering of the blog and its constituent parts derived from the SIP. Creating the DIP version might involve some form of rendering and exporting of file formats.	*Content data objects*
Discovery metadata	*Representation Information*
Rights metadata - for use by the consumers [See Deliverable D3.3.]	*Representation Information*

6.1.5 Actors

According to Brian Lavoie's study, OAIS uses Actors (also called "entities") to help define the information / preservation environment. "The OAIS environment is derived from the interaction of four entities: producers, consumers, management and the archive itself. Producers supply the information that the archive preserves. Consumers use the preserved information. A special class of consumers is the Designated Community--the subset of consumers who are expected to understand the archived information. Management is the entity responsible for establishing the broad policy objectives of the archive (e.g., determining what types of information are to be archived, identifying funding sources, etc.). The management entity does not include the day-today administration of the archive; this task is performed by a functional entity within the archive itself."

Below, we demonstrate that Producers and Consumers have already been identified in the deliverable D4.1. We also make observations about the Management and Administration entities.

Producer

OAIS: "The role played by those persons, or client systems, who provide the information to be preserved. This can include other OAISs or internal OAIS persons or systems."

The Deliverable D4.1 has identified the following producers within the stakeholder groups:

Content Providers are people or organisations which maintain one or more blogs and, hence, produce blog content that can or should be preserved in the archive. Content providers are owners of their contents and decide whether they wish to contribute their content to a preservation system or not. Therefore, it is crucial to address their needs. For content providers, we differentiate between individual blog authors and organisations which can have one or more members who blog for them, e.g. business and corporate blogs.

Individual Blog Authors. Individual blog authors maintain their own blog. Thereby, maintaining means creation of blog posts, answering comments, designing the layout of the blog, etc. Blog authors may also interpret themselves as individual authors even if they maintain their blog in connection with an organisation. It is essential to know what individual blog authors need and expect. Therefore, special emphasis was put on the examination of how blog authors currently behave and what they think or expect by a blog archive.

Organisations can serve as content providers if they maintain their own corporate blogs. Organisations with their own blogs vary from public organisations like libraries and universities to businesses. In order to be considered as content providers, they should all have in common that they are allowed to publish and distribute their blog content, and that they have an interest in its long-term preservation. Their needs have to be considered to support their organisational purposes of preservation and, thus, to increase the probability of contribution.

Management

OAIS: "**Management** is the role played by those who set overall OAIS policy as one component in a broader policy domain. In other words, Management control of the
OAIS is only one of Management's responsibilities. Management is not involved in
day-to-day archive operations. The responsibility of managing the OAIS on a day-today basis is included within the OAIS in an administrative functional entity.

Management in BlogForever may consider the following actions as part of its remit:

- Provide the BlogForever charter and scope
- Manage the source of funding
- Provide guidelines for use of resources
- Conduct review processes
- Determine or endorse pricing policies
- Support BlogForever by establishing procedures and policies, e.g. draft requirements about blogs submitted to BlogForever

Deliverable D4.1 has not yet identified a "stakeholder" that fits the management role. However, it is possible the "Manager" of BlogForever might be a *potential BlogForever company*. If BlogForever needs to identify a Management entity, this might be defined as part of the activities arising from WP6. The intention is to operate a Business Model. The charter and scope of BlogForever could come out of this process as more stakeholders are identified and defined.

Another possibility is that the BlogForever user base itself will have a stake in the Management entity, particularly with regard to sharing in the development of policies, procedures, and review mechanisms. This notion is supported by the trend of current thinking about the project, which sees BlogForever as a collaborative, community-owned platform in which all of its beneficiaries,

contributors and stakeholders have a share. In this scenario we are describing a user-driven set of policies that, over time, could grow BlogForever into a social service that will manage and regulate itself.

Administration

OAIS: "Administration is the OAIS entity that contains the services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis."

For BlogForever, the specific tasks performed by the administrators include:

1. Operation of an archive and preservation service.
2. Working to written policies and procedures for all services and functions.
3. Provision of reports on aspects of the repository.
4. Maintenance of hardware and software.

It is very likely that many of these tasks will be performed by the staff at CERN/Invenio, who are responsible for digital repository component design and administration of the final BlogForever platform.

The deliverable D4.1 has also identified the following entities, as users / stakeholders who may have an administrator role of some sort.

Next to the people who contribute to the archive or who utilise the archive, we consider **administrators** as another important stakeholder group for requirement identification. Administrators (admins) maintain installed software and will probably be responsible for a stable and robust operation of the preservation system. Thus, administrators have a different perspective on the requirements of the software and may emphasise more on technical issues, e.g. scalability. Additionally, admins can be more informative with regard to benchmarking data of current usage. Especially **admins of blog hosts** could provide valuable data about the current blogging landscape. Therefore, it is further distinguished between admins of blog hosts and admins of organisations. Thereby, the focus for the latter is on these organisations that would probably run a blog archive. We identified libraries that preserve digital information and businesses that process social media as relevant organisations.

Consumer

OAIS: "Consumer is the role played by those persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail. This can include other OAISs, as well as internal OAIS persons or systems."

Deliverable D4.1 has identified the following consumers:

Content retrievers are people or organisations which have an interest in the content stored in a blog archive and, therefore, they like to search, read, export, etc. that content. The purpose of their interest can vary broadly. They can be divided into individual blog readers, libraries, businesses, and researchers.

Individual blog readers are people who already read blogs for various reasons, e.g. family, hobbies, professional. A blog reader may also be interested in a blog archive because he/she could find blogs that he/she has read in the past but which are not available anymore. In the future, a blog reader could also be interested in the blog posts at a specific point in time, e.g. his birthday, a scientific breakthrough date, etc. Additionally, an archive could provide special functionalities that go beyond the single blog, e.g. visualize the network of blogs and recommend similar blogs. Thus,

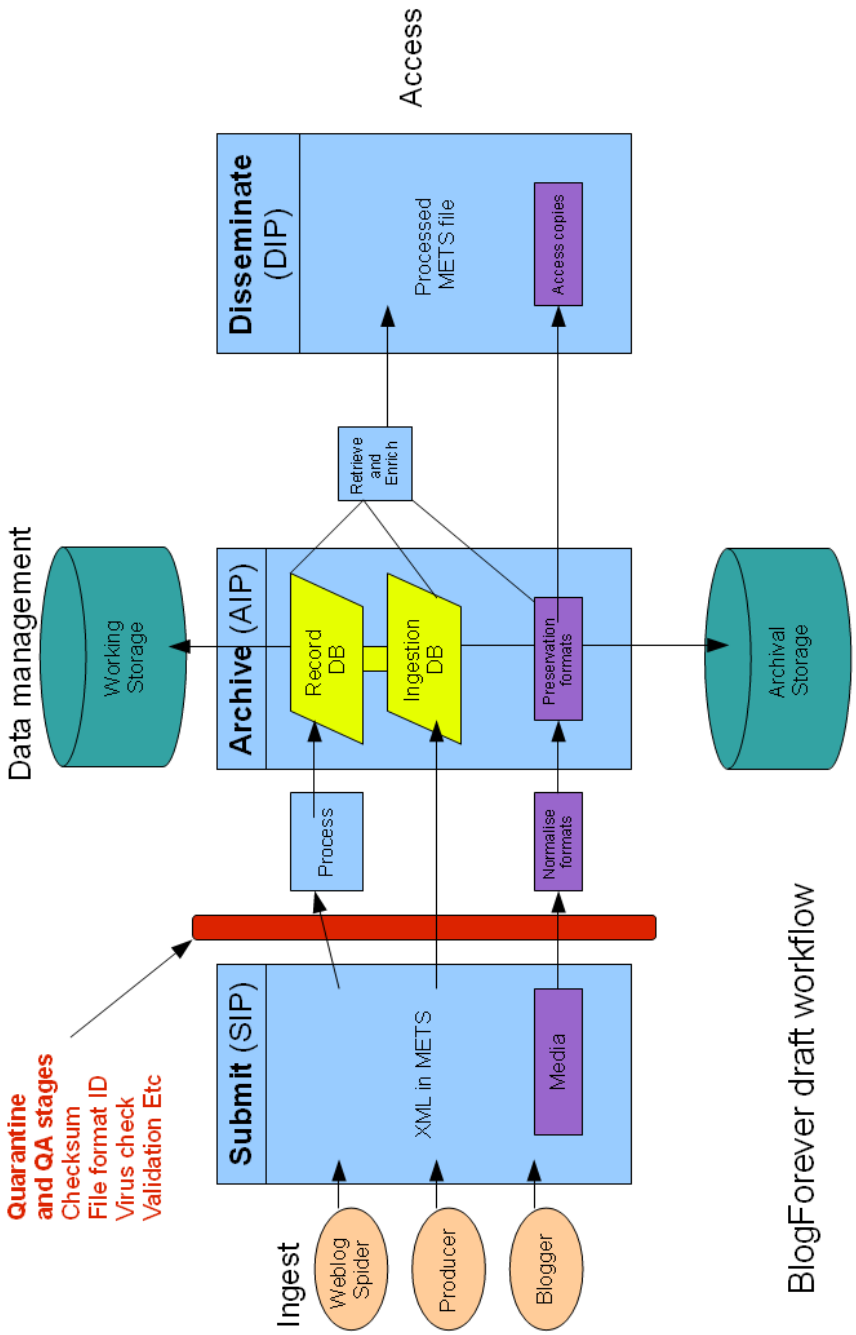
individual blog readers are an important stakeholder group from the perspective of the consumption of archived content.

In contrast, **libraries** operate more as a gatekeeper for individual retrievers. They provide access to various kinds of information sources, e.g. books, journals, movies, etc. Thereby, the access includes value added services like selecting and sorting the sources as well as adding metadata. However, libraries in their role as a gatekeeper often do not keep the content themselves, especially in the case of digital resources. Instead, they manage the references to various sources (e.g. literature databases) and if the user would like to retrieve the concrete resource, the library forwards it to the user or retrieves and delivers the resource. Libraries, in their role as gatekeepers, are very important for the adoption of the blog archive. They may have special needs for integration and access.

Businesses also offer value added services based on the available information. But contrary to libraries, they are normally more interested in processing the information to provide a unique selling proposition (USP) to their customers. Such USP could be the detection of trends or sentiments in the business field of the customers. Therefore, they collect or access available information from various sources. A real-time archive of blogs may be an interesting alternative to capturing information themselves. Thus, businesses are a promising stakeholder when business models of blog archives are considered because businesses would possibly pay for special access to archived information. Therefore, the needs and demands of businesses are also considered.

A special group of individual content retrievers is the group of **researchers**. Research on blogs can be conducted for various purposes, e.g. observation of social behaviour, inquiry of historical developments and examination of communication behaviour... However, researchers need “good” data for their research. Depending on whether they do qualitative or quantitative analysis, criteria for “good” data could be the amount of data, how representative they are, or if the author could be identified. The impact of a blog archive will increase enormously for scientific purposes if researchers' requirements are considered carefully.

6.1.6 Overview Repository Diagram



6.1.7 Conclusions

The report concludes that the BlogForever platform can and will perform digital preservation as it is understood and defined within the conceptual framework of the OAIS model. In summary:

1. A large majority of the OAIS functions and requirements are already in place at CERN
2. The prototype workflow being built by Invenio maps very closely to the core OAIS functions of Ingest, Data Management, Archival Storage and Access

3. The core repository workflow functions can be enhanced with some very simple interventions at the ingest stage, which are suggested in Sections 4 and 5 of this report
4. The planned combination of spider crawls, delivery of SIPS in XML, addition of metadata schema, database storage and delivery mechanisms produces OAIS-compliant information packages
5. BlogForever has clearly defined Actors that fit the Producers and Consumers roles in OAIS

The following areas have been identified where BlogForever is not quite an exact match for OAIS. However, they are not so critical as to cause much concern, and in our view will be addressed easily. None of them impact on the core digital preservation workflow.

1. Parts of the Administration function have not yet been fully developed at Invenio
2. There are some ambiguities regarding ownership and maintenance (post-project completion) of the Preservation Planning function
3. BlogForever does not have a clearly defined Management entity

Other details which may need clarification and assurances as the project proceeds:

- How digital objects other than XML (i.e. media attachments to blogs) will be captured and preserved, particularly if they need to be migrated. See 5.1.4.
- How managed storage and backing up will be performed. See 5.2.2.

6.2 Repository Risk: DRAMBORA for Weblogs

The Digital Repository Audit Method Based on Risk Assessment[210] (DRAMBORA) was developed by jointly by the Digital Curation Centre[211] (DCC) and DigitalPreservationEurope[212] (DPE). The approach is a framework for the self-assessment, encouraging repository awareness of their objectives, activities, and assets, and supporting them to identify and estimate risks implicit within their organisation. The assessment takes place using the following general steps[213]:

- Defining functions of the repository.
- Identifying the activities and assets associated with repository functions.
- Assessing the risks that might be associated with the activities and assets.
- Calculating risk impacts.
- Planning how the risks might be treated
- Reporting on the self-audit

The assessment is recommended within three contexts: for the validation of an existing repository, as a means of preparing a repository for an external audit, and for the identification of gaps in anticipation of a future repository in development.

This section presents a description of the BlogForever project WP3 effort to apply the method to the weblog repository context. It was deemed as a worthwhile intellectual exercise that would make the weaknesses and strengths of a repository concrete and explicit. This would be in contrast to the high level conceptual framework of OAIS. The exercise was carried out within the anticipatory context of a future repository. The task proved to be more difficult than first envisioned. Part of the difficulty originated from: a) the abstract nature of the task due to the fact that the assessment was anticipatory (that is, there was yet no existing repository), b) the abstract nature of the organisational objectives due to the fact that the project would not be taking on the actual

[210] http://www.repositoryaudit.eu/about/
[211] http://www.dcc.ac.uk
[212] http://www.digitalpreservationeurope.eu/
[213] http://www.repositoryaudit.eu/objectives/

responsibility ourselves of building, running, and sustaining a weblog repository (that is, the final outcome of the project would consist of a tested prototype repository software and preservation planning guidelines for future organisations wishing to archive weblogs).

As a first step, members of the project were asked to assign their assigned tasks within the project to one of the functions from the DRAMBORA functional classes (Figure 5-1).

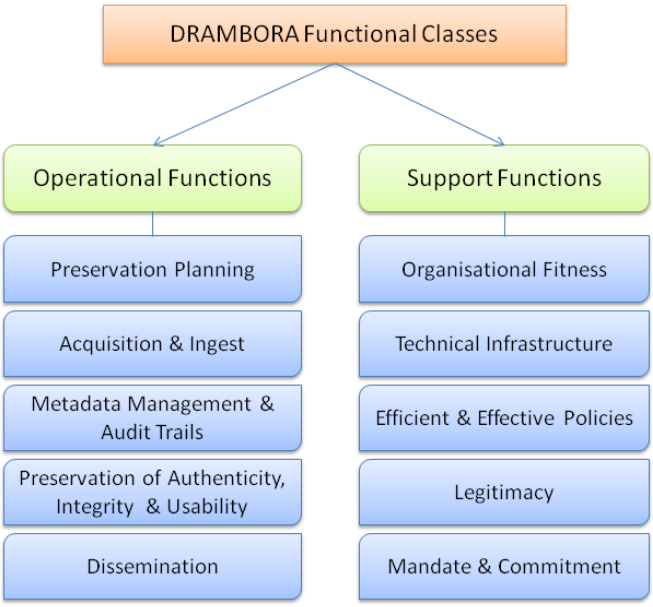


Figure 5-6.2-1 Repository Functions Considered as Part of DRAMBORA

As an initial objective the project team wanted to make sure that the main operational functions of the repository were covered by the project activities involved in repository implementation and preservation policy development. The diagrams in the following figures are intended to show the range of activities within the project that cover the five main repository function (left hand column of Figure 5-6.2-1). The activities are indicated in circles with orange coloured border line, along with their corresponding work package task number.

The activities, in some cases, have been simplified so that where design will be followed by implementation (for example, this is the case for Tasks 4.2, 4.4), only the design phase has been indicated. In Table

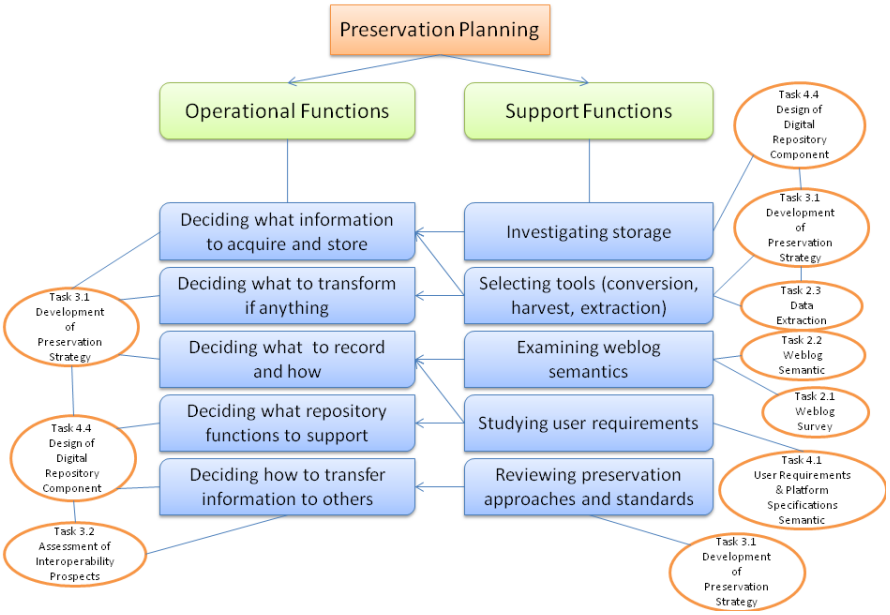


Figure 5-6.2-2 Mapping Project Activities to Preservation Planning Components

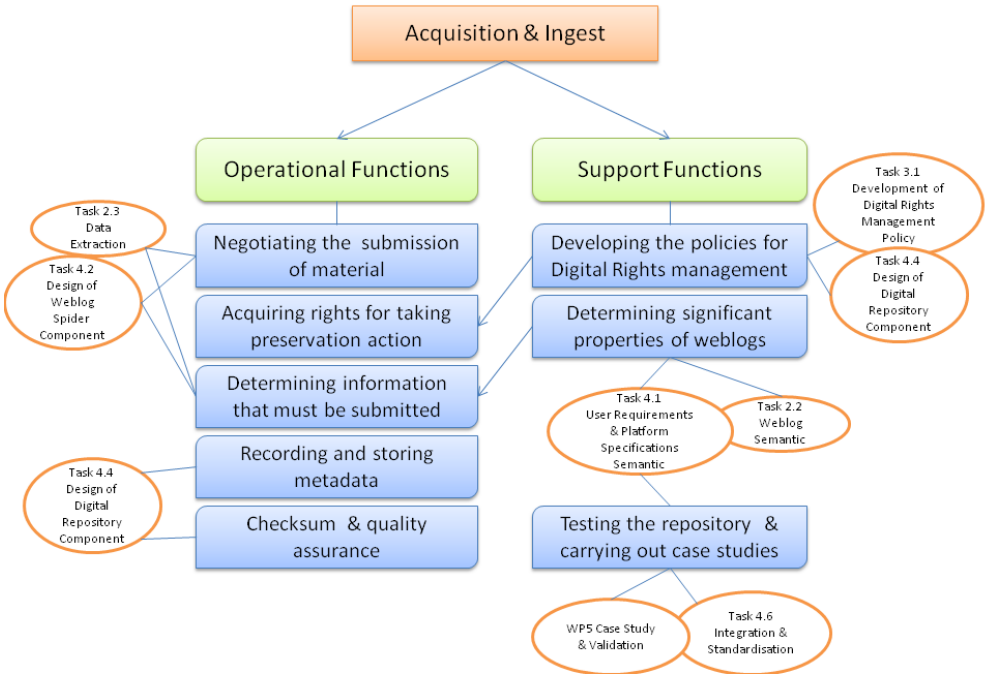


Figure 5-6.2-3 Mapping Project Activities to Acquisition and Ingest Components

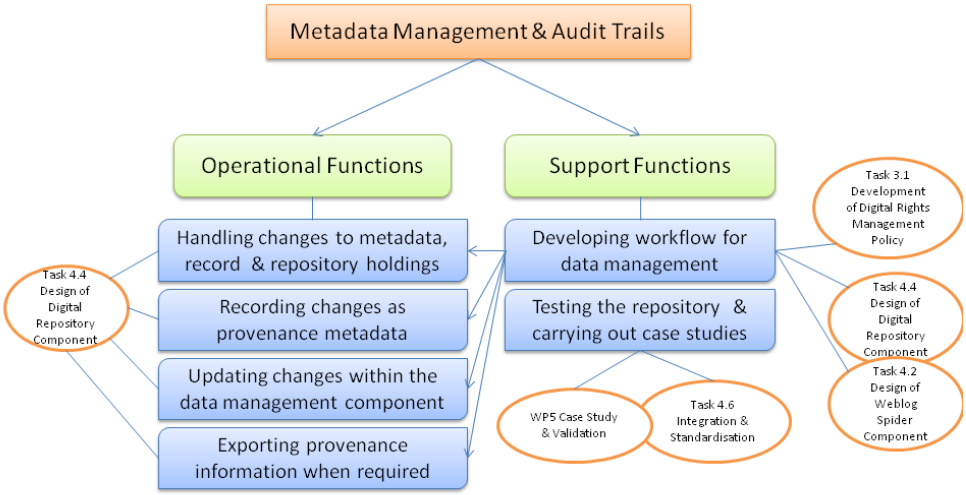


Figure 5-6.2-4 Mapping Project Activities to Metadata Management & Audit Trail Components

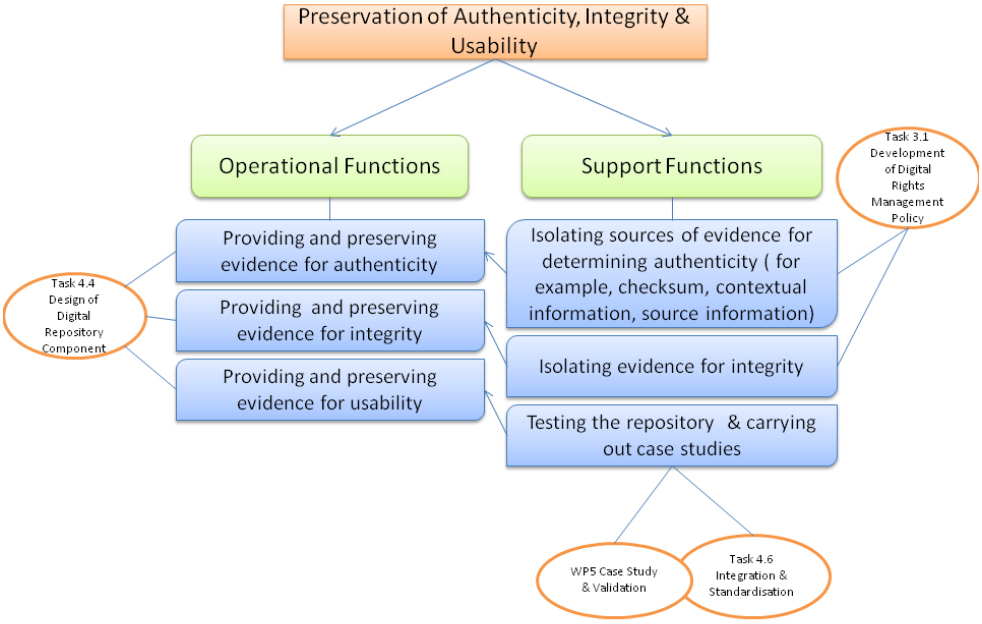


Figure 5-6.2-5 Mapping Project Activities to Preservation of Authenticity, Integrity and Usability Component

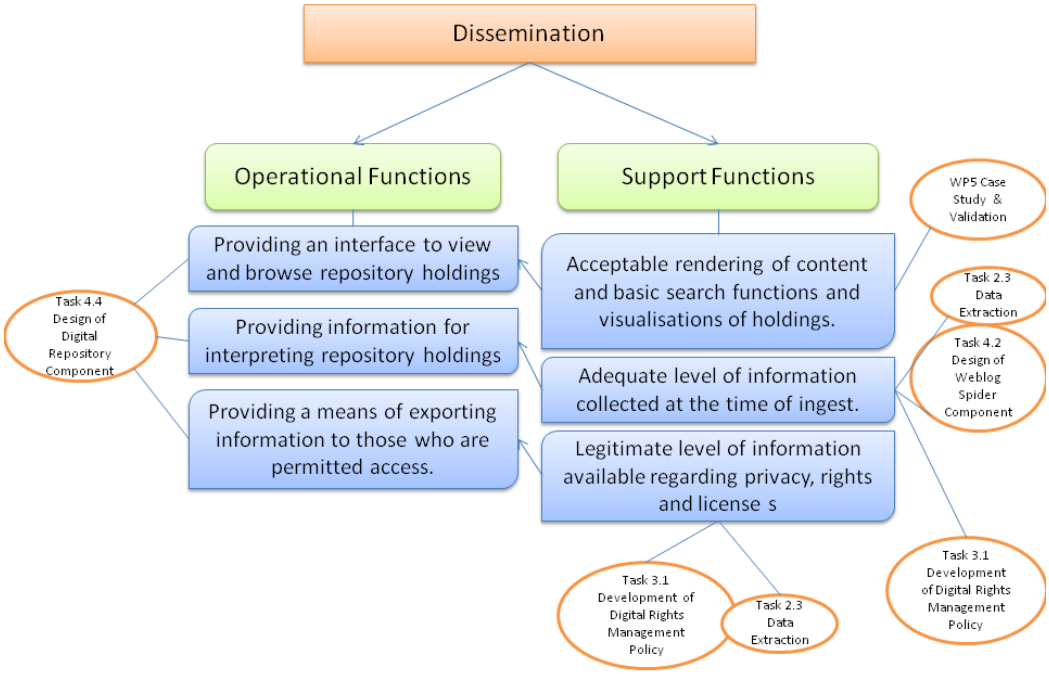


Figure 5-6.2-6 Mapping Project Activities to Dissemination Components

The activities named in the figures above will be explicitly supported through the project reports listed in the last column of Table 6.2-1. The assets in bold are intended to be those resulting from the activity in the middle column, while those not in bold are intended to be the assets that support the activities.

Table 6.2-1 Assets associated to repository activities.

Function	Activity	Assets
Preservation Planning	WP3 Development of Preservation Strategy. WP4 Design & Implementation of Digital Repository Component. WP3 Assessment of Interoperability Prospects.	D2.1 Weblog Survey Report D2.2 Report on Weblog Data model D2.3 Weblog Ontologies D4.1 User Requirements and Platform Specification **D3.1 Preservation Strategy Report** **D4.4 Digital Repository Component Design**
Acquisition & Ingest	WP2 Data Extraction WP4 Implementation of Weblog Spider Component WP4 Implementation of Digital Repository Component	D2.4 Spider Prototype D2.5 Spam Filtering Report D2.6 Data Extraction Methodology D3.1 Preservation Strategy Report D5.5 Case studies comparative analysis & conclusions **D4.5 Implementation of Repository Component Design**
Metadata Management & Audit Trail	WP4 Implementation & Standardisation of Digital Repository Component	D3.1 Preservation Strategy Report D3.3 Digital Rights Management Policy **D5.5 Case studies comparative analysis & conclusions**

Function	Activity	Assets
		D4.5 Implementation of Repository Component Design
Preservation of Authenticity, Integrity, and Usability	WP4 Implementation & Standardisation of Digital Repository Component	D3.1 Preservation Strategy Report D3.3 Digital Rights Management Policy D5.5 Case studies comparative analysis & conclusions **D4.7 Final Weblog Digital Repository**
Dissemination	WP4 Final BlogForever Platform	D3.3 Digital Rights Management Policy D3.2 Assessment of Interoperability D5.5 Case studies comparative analysis & conclusions **D4.8 Final BlogForever Platform**

The repository's preservation strategy depends on the understanding of the semantics and the significant properties that add value to weblogs. Likewise, the repository system design and implementation depends on the development of mechanisms that support the preservation of the identified semantics and properties. These properties, however, are fluid: for example, depending on the purpose of the community or organisation preserving the weblogs, the most significant properties can differ noticeably.

The lack of an organisational perspective (due to the fact that we ourselves are not building a weblog repository – only providing the means for others to do so), posed some difficulty in formulating an approach to identifying and calculating risk impacts. Instead, we discuss the risks involved in the choices we make with respect to tasks above in Chapter 7 and their impact associated to sustaining the communities surrounding weblogs.

7 BlogForever Preservation Strategy

The discussion in this chapter integrates the results from Chapters 3, 4, 5, 6 to present an outline of the recommended BlogForever strategy for the preservation of weblogs. In the current proposal we have discussed common components and associated objectives and formats that appear in blogs (Section 3.1 and 3.2), and derived significant properties from them (Sections 3.3 and 3.4), We risks of information loss (Section 4.2) and how a better characterisation of data complexity that could help us to assess risks (Sections 4.3) and support designated weblogging communities (Sections 4.3 and 4.4). The recommendation in this report is to express the characterisation of weblogs developed in Chapters 3 and 4 using the metadata schemas recommended in Sections 5.3 and 5.4 this will be wrapped in METS[214] (see Section 5.5). This process can is mapped to a sub process of the Ingest process described in Section 6.1, anticipated to take place during or after the quality assurance and description process (preservation service recommendations 2 and 3, Section 6.1.2). The broader set of preservation service recommendations that will map the workflow of the repository to OAIS-like functions is described in Section 6.1.

Here we would like to expand on three aspects of the repository that will be added to the recommendations in Chapter 6. These are added to enhance the robustness of the repertory preservation functions, e.g. they are intended to mitigate some of the risks that were discussed in Section 4.2. More specifically, we conclude our recommendation for a weblog preservation strategy with detailed recommendations with respect to storage (Section 7.1), blog characterisation (Section 7.2), and end-user repository features (Section 7.3).

7.1 Recommendations for Storage: Keeping More than What Is Perceived to Be Valuable Now

By plucking blogs out of their natural habitat and storing them away in a repository, we take on the role of crime scene investigators collecting evidence from the scene of a crime. Without extreme caution, investigators are apt to run the risk of contaminating vital evidence that would have helped to solve the crime.

Nevertheless, current curators of web information are only too happy to dismantle the web as it stands: they envision what might be disseminated as the end product of the repository, augment it with descriptive and administrative metadata, and collect what might serve as adequate proxies for satisfactory values for the predefined set of metadata elements.

As a long-term preservation principle, this way of curating and managing information has some immediately noticeable limitations: a) the information loss resulting from the evidence that we did not collect is inestimable, b) the discovery of novel or unexpected connections between different types of information becomes increasingly difficult, and, perhaps most importantly, c) we do not know what information we will be accessing in the future and how we will be accessing the information we collect in the future.

For example, there are logs of errors and processes as well as technical clues that provide insight into what kind of information and/or software was required and accessed in rendering selected digital objects. This information is rarely collected, if ever. Likewise there may be information that leads to insight regarding community interaction that does not surface on the basis of information only (see Chapter 4). In many cases these surface only as part of syntactic or pragmatic signatures and/or traces of information: for instance, new languages such as HTML5 emerge on the basis of

[214] http://www.loc.gov/standards/mets/

web authoring statistics[215] and approaches to webpage structure/design search are developed on the basis of web technology usage statistics[216]

The International Internet Preservation Consortium (IIPC) showed how the way we are accessing information is changing and how we need to provide better methods to the users in accessing archives[217]. The problem is that the novel access methods are dependent on machine learning and statistical analysis strategies that rely on the availability of sufficient *representative* data. Without the necessary data, the tool may be in place but will perform poorly and reflect on the trustworthiness of the archive.

Machine learning, pattern recognition and statistical analysis has become an ubiquitous approach to data access both in the sciences and in the newspapers. Archives that do not preserve data to meet the requirements of these technologies especially with the current emphasis on big datasets (such as that collected within web archives) will quickly become obsolete and unusable.

Figure 7.1-1 Storage workflow diagram.

Even if we are enlightened enough to be able to collect the correct information to be stored for posterity, it is doubtful that we can indefinitely maintain a system that will cope with the increasing volume of information and complexity of preservation processes while sustaining an acceptable level of accuracy. Even now the accuracy of format identification, object characterisation and format validation tools are variable depending upon the selected tool.

In light of these observations we suggest an approach to storage that can allow for re-construction and re-interpretation of information. In Figure 7.1-1, we have displayed three of the core components of information packages in an archive. We suggest that, while, the archival information package and dissemination package might follow the recommendations of traditional repositories, including encapsulated metadata and representation information, the submission information package be replaced by a stored information component consisting of the packets exactly as they are received as a response to the http request issued by a web spider that aims to retrieve the entire webpage.

The information retrieved by the web spider must be stored in a robust format. There are seven core attributes for what might constitute a robust storage format for preservation (Kim and Ross 2011b). These are:

1. "Completeness of data: the format should preserve data as closely as possible to a sector-by-sector copy of the raw data on a system disk, for example, inclusive of file structure, dependencies, and process history."
2. "Recoverability of data: the format should support the recovery of data wherever possible, e.g. one corrupted file or sector, if possible, should not pose serious problems in recovering other files and sectors in the archive."
3. "Robust support for data validation: for instance, it is recommended that the format should provide a method for processing piecewise hash codes (for arbitrary bitstreams).
4. Scalability of data management processes: for example, no limitations should be placed on input/output size and/or media types and random access should be possible.

[215] https://developers.google.com/webmasters/state-of-the-web/
[216] http://dev.opera.com/articles/view/mama/
[217] http://netpreserve.org/publications/2011_06_IIPC_WebArchives-TheFutures.pdf

5. Transparency: the format specification should be a publicly published open standard and source allowing modification, distribution and the tracing of accountability for preservation purposes.
6. Flexibility of embedding metadata: ideally, the format should allow the inclusion of metadata of arbitrary type, schema and length.
7. Flexibility of handling data: it is recommended that the format should be able to handle data objects in its entirety or in small portions, on any media types and from any source (e.g. streamed data as well as stored data).

At the time of the study, we found the digital forensics format, Advanced Forensics Format (aff), close to meeting these criteria. We recommend that the best format be investigated further to reach a conclusive decision.

7.2 Taking advantage of diversity: looking for digital finger prints

The trend in archiving practices has been to formally define characteristics of archived records or information that preserve the evidence necessary for the evaluation of their authenticity. While this could be no different in the case of web archiving, the notion of what constitutes evidence for authenticity may need further consideration.

In the analogue context, there are well understood *fingerprints* of expression: for example, handwriting, styles and conventions that manifest themselves in a visible form on paper, the choice of medium, say, for instance, clay, or other tactile material. In the digital world, these are hidden, and often emerge as choices in technology, medium, formats and object types. For example, the use of PHP scripts that generate images from LaTeX is evidence that supports the conjecture that a blog is written by a mathematician or scientist.

The focus that has been laid on web content, has led the web archives, thus far, to spend their best efforts in preserving what is viewed as the immediate semantic information content of the web page. This could lead to the elimination of valuable evidential support for evaluating the authenticity of information and associated accountability by disabling the possibility of tracing the information to its source and historical context.

The *uniformisation* of the technical aspect of the content at the time of archiving also causes loss of historical perspective on the technology itself. In the long-term, the history of technological change is bound to form a relevant part of our cultural legacy. While it is understandable that the repository managers might create a uniform structure for the access copy of the archive to facilitate easy management and renderability of the record, in BlogForever, we propose that the technological characterisation of blogging communities be available as part of the provenance contextual information.

The current workflow of the repository suggests that the URL of a desired blog will be submitted to the repository. This will be passed to the weblog spider which will harvest the requested pages and return it to the repository (the process labelled 1 in Figure 7.2-1). The Weblog spider will also gather some metadata in the data extraction process and return it to the repository (the process labelled 2 in Figure 7.2-1). These may be combined in the implementation as one process but it has been divided here to make it explicit. Once the harvested page and extracted metadata arrives at the repository it is recommended that the weblog data go through a second stage of characterisation followed by an update of the metadata (the process labelled 3 in Figure 7.2-1). This second stage will be in two parts. The first of these consists of format identification and characterisation, to extract the technical metadata recommended in Section 5.4.1 for digital object types embedded in the weblog. The second of these consists of extracting additional contextual metadata such as the characteristics that are indicative of the blogging community as described in Section 4.4.

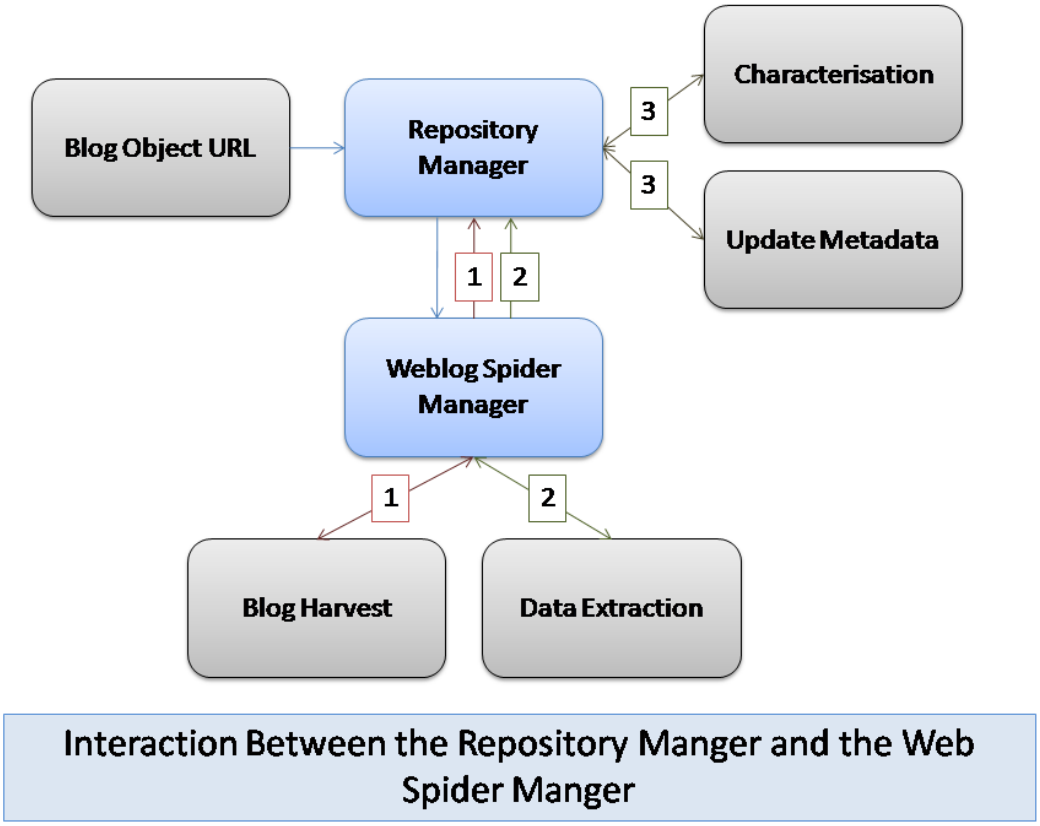


Figure 7.2-1 1 Repository features recommended for characterising format specific technical metadata and community specific provenance information.

The extracted contextual information should be mapped to the components of the data model included in the four record types to serve as contextual and provenance information for the target weblog. The two processes for the refining metadata, that is, administrative metadata for the weblogs arriving into the repository, is described in Figure 7.2-2.

Ultimately these processes may prove too intensive to perform for all material coming into the repository, but the viability of doing this should be thoroughly evaluated within the scope of the project resources.

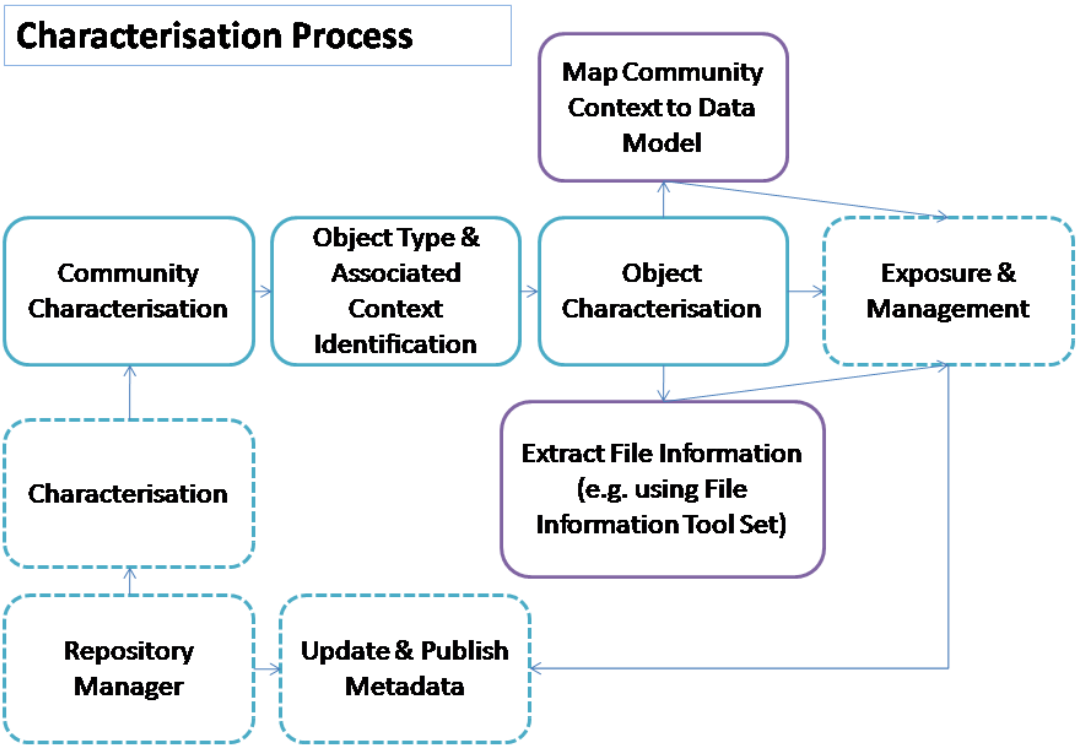


Figure 7.2-2 Refining metadata for a weblog after it has been harvested and transmitted to the repository.

7.3 Redirecting expert attention: getting the community involved

One way to sustain digital information is to keep it in use. This has three effects:

1) Any problems with access will be detected early before all information relevant to its recovery is completely lost.
2) The fact that it is being used is likely to imply that it is of value to someone in the community, that is in someone's interest to preserve it.
3) In relation to 2), there is community support for finding solutions preservation and information access problems.

Currently there is much discussion of crowd sourcing as a means of gathering information. Initially this took the form of channels such as Wikipedia[218] that provide collaboratively refined information, and Amazon Mechanical Turk[219] that functions as a crowd sourcing tool. However, the horizon for crowd sourcing is quickly expanding to include crowd sourcing for specialised information[220]. The anti-bot service, reCaptcha[221] has been used as a crowd sourcing device to improve image recognition tasks.

[218] http://www.wikipedia.org/
[219] https://www.mturk.com/mturk/welcome
[220] http://dailycrowdsource.com/crowdsourcing/articles/microvolunteering/436-think-you-can-classify-a-galaxy
[221] http://www.google.com/recaptcha

There is no reason why such crowd sourcing devices cannot be incorporated into the repository to support preservation. In fact, it has been observed that "cooperation" might be a feature that naturally evolves within society[222]. Given that weblogs are created by people who like to interact online, the potential for cooperation to create a better repository may even be better.

It is the recommendation of this proposal that some of the features (examples are displayed in Figure 7.3-1) that would result in the improvement of the repository and preservation support be developed as end-user functionalities.

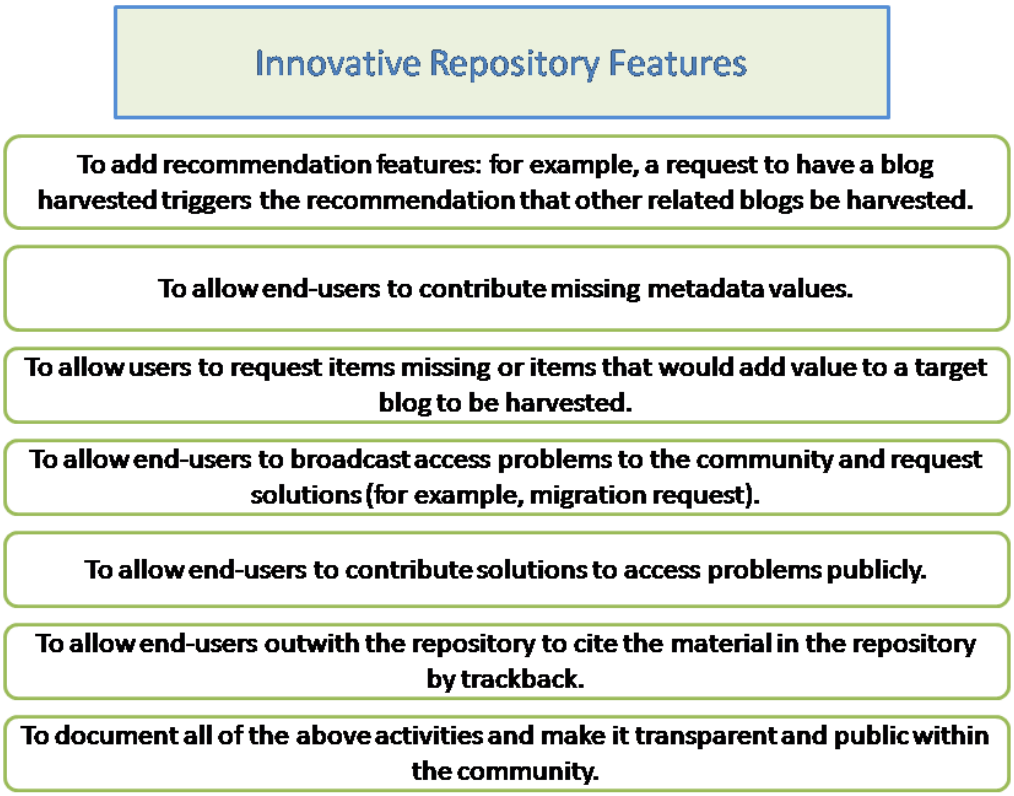


Figure 7.3-1 End-user repository features that would improve repository quality and support preservation of weblogs.

Including innovative features that serve to refine the quality of the repository (e.g. users are allowed to provide missing metadata), add value to material already in the repository (e.g. trackback functionality; users are allowed to request additional material to be connected to an existing blog) and improve preservation activities (e.g. migration on demand) is imperative to generate a solid business model that would attract future adopters of the BlogForever platform.

[222] http://www.ped.fas.harvard.edu/people/faculty/publications_nowak/Nowak_Science06.pdf

8 Conclusions

In this report we have discussed topics that have led to the development of a recommended strategy for weblog preservation. In this chapter we conclude the report by summarising what this report contributes to the current research landscape (Section 8.1), what we have learned through the process (Section 8.2) and directions that might be considered for future work (Section 8.3).

8.1 Contributions of This Report and How to Take it Forward

In this report we discussed why we might want to preserve weblogs, what properties of the blogs we would want to preserve and how we might support their preservation in a repository frame work. The work consisted of reviewing the

- Potential values that have been observed in relation to weblogs.
- Contributions from BlogForever: D2.1 Weblog Survey, BlogForever: D2.2 Weblog Data Model, BlogForever: D4.1 user Requirements and Platform Specifications Report.
- Previous work in the preservation of digital objects.
- Risk of information loss with respect to weblogs.
- Characteristics of currently active weblogs.
- Widely adopted metadata schemas and encoding standards.
- The reference model for an Open Archival Information System (OAIS).
- The Digital Repository Audit Method Based On Risk Assessment (DRAMBORA).

This resulted in

- A description of weblog properties that need to be preserved to meet user requirements.
- A description of properties that need to be preserved in relation to objects embedded within weblogs.
- A characterisation of weblog complexity.
- A study of the relationship between weblog complexity and weblog community.
- A practical repository work flow that contextualises the BlogForever repository with respect to the OAIS.
- A mapping between project objectives and DRAMBORA.
- Recommended practices in adopting metadata schemas for recording the properties of weblogs and their associated components.
- Recommendations on how the metadata will be encoded and shared using the Metadata Encoding and Transmission Standard (METS).
- Recommendation regarding the storage of weblog information.
- Detailed characterisation processes to refine administrative metadata associated with weblogs.
- Innovative repository features that, if implemented would enhance the repository quality, value, and support for preservation.

The work was summarised and integrated as a preservation strategy in Chapter 7. Clear recommendations for repository features have been made in Chapters 5 (in terms of recommended metadata schemas and encoding standards), Chapter 6 (in relation to preservation service recommendation), and Chapter 7 (with respect to archival storage formats, refinement administrative metadata extraction, and innovative repository features that enhance preservation activities). The recommended features will require further discussion with BlogForever WP4 before it can be finalised formally within subsequent deliverables, such as deliverable D4.4 Design of the Repository Component and future refinements.

8.2 What We Learned

The results of the work here show that weblogs are complex evolving objects. We have discussed the most prominent features that have been observable within the limits of the resources and time available within this project. Instead of focusing on the full array of complex object types that only surfaced rarely in the datasets that we were examining, in this project, we have initially opted to focus on the interactive aspects of weblogs such as interconnections between components and how these change over time.

However, it is clear that the types of objects embedded in webpages are increasingly becoming complex with animation features and layered images (e.g. see the web page here: http://jessandruss.us/). This has brought to light that

- There is an urgent need to develop scalable approaches to implement increasingly complex preservation processes within the repository.

We need to think forward and not limit ourselves to solutions that depend on the number of servers or distributed computing only. These have hard limits depending on the required process. And, also, they are solutions that work well on homogeneous collection with to deal with volume. They do not necessary work efficiently when the collection is heterogeneous and complex and structure.

The conclusion of this work is that we are not yet ready to apply PLANET- style experimentation on weblogs to examine adequate support for the complexities.

On the other hand there are other ways of circumventing problems of scalability such as creating focused collections based on selected blogging communities and crowd sourcing (some repository features that would support this have been suggested in the previous chapter.

The work in this report has to be taken further. We really need to understand better what goes on in a digital object and how we can characterise them without having to recreate a model for every situation. In the following section we have outlined some suggestions for future work that could lead to fruitful way to address these challenges.

8.3 Future Work

In Figure 8.3-1, we have presented some areas of research that could take the current research in weblog preservation forward.

There has been a lot of work on the formats that best preserve digital information of a selected type. On the other hand there has been very little work to evaluate comprehensively formats on the basis of their functionalities. For example, as far as we know, storage container formats are not as extensively investigated within the literature. A bench-marked list of recommended format based on a comprehensive study with respect to different types of content would be useful.

The data analysis in Chapter 4, revealed some interesting relationships between the blogging community and technical aspects of the webpage. This needs to be extended to study other hidden relationships, not only between the blogging community and the technical aspects of the pages, but relationships between different elements. This could bring to light “significant properties” relevant to preservation activities.

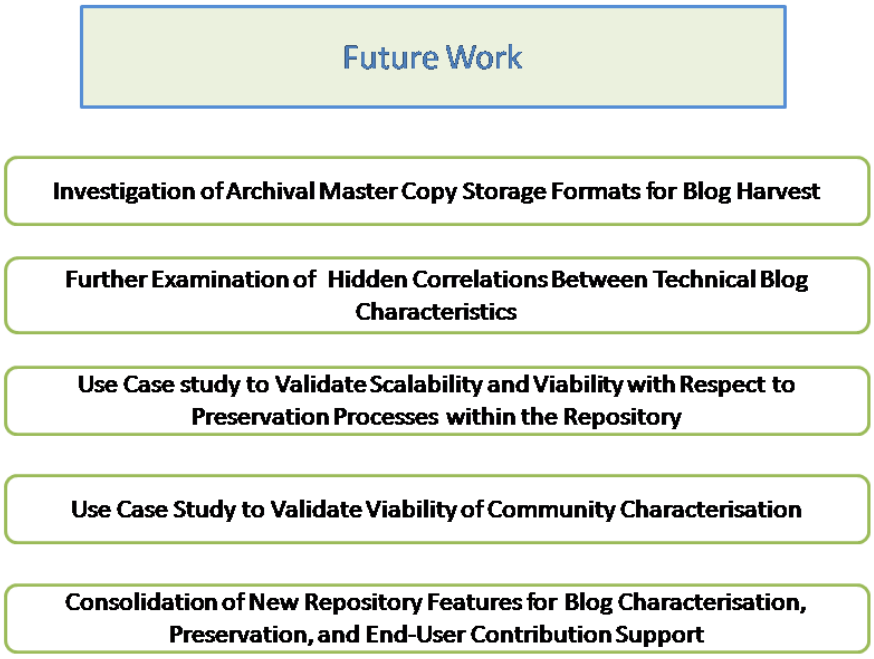


Figure 8.3-1 Suggested future work for the preservation of weblogs.

In the current work we found that there are serious scalability issues involved in using file characterisation tool in the "big data" environment. This has also been noted elsewhere (e.g. within the SCAPE project, and experiments conducted by the IIPC), however, most studies have been conducted on categorised archived material, reducing the consideration of the complexities that were inherent in the source object. It is imperative that the demand on the processes introduced by object complexity is evaluated.

We have proposed a characterisation of weblogs on the basis of their complexity measured by the variation that exists within the collection. We have also shown evidence that this complexity is bound to the blogging community to which the blogs belong. This notion of complexity should be extended and formalised to be used in BlogForver use case studies as a means of selecting representative data and to see if it can be used as provenance information.

The future work suggested here can be taken forward as part of BlogForever only after careful consideration, discussion, and, consolidation to produce formal specification of what their implementation might entail.

9 References

Aitken, B., Helwig, P., Jackson, A., Lindley, A. Nicchiarelli, E., Ross, S. (2008) "The Planets Testbed: Science for Digital Preservation." Code4Lib Journal, Issue 3, 2008-06-23, ISSN 1940-5758. http://journal.code4lib.org/articles/83

Ambacher, B. u. a. (2007), *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)*, CRL Center for Research Libraries, Chicago, IL. Available from: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

Archivematica, release 0.6-alpha. See http://archivematica.org.

Banos, V., Stepanyan, K., Joy, M., Cristea, A. I. and Manolopoulos, Y. *Technological foundations of the current Blogosphere*. City, 2012.

Bar-Yossef, Z., Broder, A. Z., Kumar, R., and Tomkins, A. (2004) "Sic Transit Gloria Telae: Towards an Understanding of Web Decay." In *Proceedings of the 13th international conference on World Wide Web* (WWW '04). ACM, New York, NY, USA, 328-337. http://delivery.acm.org/10.1145/990000/988716/p328-baryossef.pdf (accesed 16/08/2012)

Becker, C., Kulovits, H., Rauber, A. and Hofman, H. *Plato: a service oriented decision support system for preservation planning*. ACM, City, 2008.

Berners-Lee, Tim (1998). "Cool URIs Don't Change." Online Article. http://www.w3.org/Provider/Style/URI.html (accessed 16/08/2012).

Besser, H. (1990) "Visual Access to Visual Images: The UC Berkeley Image Database Project." Library Trends, Volume 38, Number 4, 787-798.

Brügger, N. *Archiving Websites. General Considerations and Strategies*. Blackwell Publishing, City, 2011.

Caverlee, J., & Webb, S. (2008). "A large-scale study of MySpace: Observations and implications for online social networks." *Proceedings of the Second International Conference on Weblogs and Social Media.*

Chen, X. (2012) "Blog Archiving Issues: A Look at Blogs on Major Events and Popular Blogs." Internet Reference Services Quarterly. 21-33. http://dx.doi.org/10.1080/10875300903529571 (accessed 16/08/2012).

Coulon, F. (2005) "The use of Social Network Analysis in Innovation Research : A literature review." Available at http://www.druid.dk/conferences/winter2005/papers/dw2005-305.pdf (accessed 25 September 2012)

Dappert, A. and Farquhar, A. (2009) "Significance is in the eye of the stakeholder." In Proceedings of the 13th European conference on Research and advanced technology for digital libraries (ECDL'09), Maristella Agosti, Jose Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas (Eds.). Springer-Verlag, Berlin, Heidelberg, 297-308.

Digital Curation Centre (DCC), and Digital Preservation Europe (DPE) (2007), DCC and DPE *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)*, Digital Curation Centre, Edinburgh, UK. Available from: http://www.repositoryaudit.eu/download

DCC Curation Centre (2009) “DCC Methodology for Designing and Evaluating Curation and Preservation Experiments V1.1”, Digital Curation Centre Publication. http://www.era.lib.ed.ac.uk/bitstream/1842/3376/1/Kim%20TestBedMethodV1.1.pdf

Deken, J. M. Preserving Digital Libraries. *Science & Technology Libraries*, 25, 1-2 (2004/11/29 2004), 227-241.

Dillon, T., Chang, E., Hadzic, M. and Wongthongtham, P. *Differentiating conceptual modelling from data modelling, knowledge modelling and ontology modelling and a notation for ontology modelling*. Australian Computer Society, Inc., City, 2008.

DINI AG Elektronisches Publizieren (2006), *DINI-Certificate Document and Publication Services 2007 (Version 2.0)*, Deutsche Initiative für Netzwerkinformation (DINI), Göttingen, Germany. Available from: http://nbn-resolving.de/urn:nbn:de:kobv:11-10075687

Doerr, M. and Tzitzikas, Y. Information Carriers and Identification of Information Objects: An Ontological Approach. *Arxiv preprint arXiv:1201.0385*2012).

Dobratz, S. et al. (2006), *Catalogue of Criteria for Trusted Digital Repositories*, Die Deutsche Bibliothek, Frankfurt (Main), Germany. Available from: http://edoc.hu-berlin.de/series/nestor-materialien/8/PDF/8.pdf

Dobratz, S. et al. (2009), *Catalogue of Criteria for Trusted Digital Repositories*, nestor materials, Deutsche Nationalbibliothek, Frankfurt (Main), Germany. [online] Available from: http://nbn-resolving.de/urn:nbn:de:0008-2010030806

Dollar, C. (1971) "Documentation of Machine Readable Records and Research: A Historian's View," Prologue: The Jounral of the National Archives 3 (Spring, 1971), 27 - 31.

Farquhar, A. and Hockx-Yu, H. Planets: Integrated services for digital preservation. *International Journal of Digital Curation*, 21, 2 2007), 88-99.

Garden, M. Defining blog: A fool’s errand or a necessary undertaking. *Journalism*(20 September 2011 2011), 1-17.

Garrett, J., D. Waters, H. Gladney, P. Andre, H. Besser, N. Elkington, H. Gladney, M. Hedstrom, P. Hirtle, K. Hunter, R. Kelly, D. Kresh, M. Lesk, M. Levering, W. Lougee, C. Lynch, C. Mandel, S. Mooney, A. Okerson, J. Neal, S. Rosenblatt, and S. Weibe (1996). "Preserving digital information: Report of the task force on archiving of digital information" Commission on Preservation and Access and the Research Libraries Group. http://www.oclc.org/resources/research/activities/digpresstudy/final-report.pdf (accessed 24/08/2012)

Gero, J. S. Design prototypes: a knowledge representation schema for design. *AI magazine*, 11, 4 1990), 26.

Gomes, Daniel; Silva, Mário J. (2006) "Modelling Information Persistence on the Web" *Proceedings of The 6th International Conference on Web Engineering. ICWE'06*. http://xldb.di.fc.ul.pt/daniel/docs/papers/gomes06urlPersistence.pdf (accessed 16/08/2012)

Granger, S. (2000) *Emulation as a Digital Preservation Strategy*. Technical Report. Corporation for National Research Initiatives. http://www.dlib.org/dlib/october00/granger/10granger.html (accessed 24/08/2012).

Hank, C. *Blogger perspectives on digital preservation: Attributes, behaviors, and preferences*. City, 2009.

Hanson-Smith, E. (2012) Online Community of Practice. The encyclopedia of Applied Linguistics. Blackwell Publishing. DOI:10.1002/9781405198431.wbeal0883

Hedstrom, M. (1997) "Digital Preservation: a time bomb for digital libraries." *Computers and the Humanities* Volume 31, Number 3 (1997), 189-202, DOI: 10.1023/A:1000676723.

Hedstrom, M. and Lee, C. A. *Significant properties of digital objects: definitions, applications, implications*. Luxembourg: Office for Official Publications of the European Communities, City, 2002.

Herring, C. Scheidt, L.A., Bonus, S., and Wright, E. (2004) "Bridging the Gap: a genre Analysis of Weblogs." Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04), January 2008, Big Island, Hawaii, USA. http://doi.ieeecomputersociety.org/10.1109/HICSS.2004.1265271 (accessed 16/08/2012)

Hockx-Yu, H. and Knight, G. What to preserve?: significant properties of digital objects. *International Journal of Digital Curation*, 3, 1 2008), 141-153.

Hull, E., Jackson, K. and Dick, J. *Requirements engineering*. Springer-Verlag New York Inc, 2010.

Kalb, H., Kasioumis, N., García Llopis, J., Postaci, S. and Arango-Docio, S. *BlogForever: D4.1 User Requirements and Platform Specifications Report*. Technische Universität Berlin, 2011.

Kalb, H., Kim, Y. and Lazaridou, P. *BlogForever: Weblog Ontologies*. 2012.

Kenney, A. R. And Personius, L. K. (1992) The Cornell/Xerox Commission on Preservation and Access Joint Study in Digital Preservation. Report: Phase 1 (January 1990-December 1991). Digital Capture, Paper Facsimiles, and Network Access. Potentially available online: http://eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED352040&ERICExtSearch_SearchType_0=no&accno=ED352040

Kim, Y. And Ross, S. (2011a) "Preserving change: observations on weblog preservation." In: International Conference on Preservation of Digital Objects (iPres 2011), 1-3 Nov 2011, Singapore.

Kim, Y., and Ross, S. (2011b) "Digital forensics formats: seeking a digital preservation storage format for web archiving." In: *International Digital Curation Confenrence (IDCC 2011)*, 5-7 December 2011, Bristol, UK.

Knight, G. (2005) SHERPA-DP OAIS Report: An OAIS compliant model for Disaggregated services, Version 1.1, 2005.

Knight, G. and Pennock, M. (2009) Data without meaning: Establishing the significant properties of digital research. *International Journal of Digital Curation*, 4, 1 2009), 159-174.

Lave, J., & Wenger, E. (1990). *Situated learning: Legitimate peripheral participation.* Cambridge, England: Cambridge University Press.

Lomborg, S. Navigating the blogosphere: Towards a genre-based typology of weblogs. *First Monday*, 14, 5 2009).

Long, A. S. (2009) “Long Term Preservation of Web Archives – Experimenting with Migration and Emulation Methodologies.” IIPC project to evaluate emulation and migration as long-term preservation solutions for web archives.

Masanès, J. Web Archiving: issues and methods. *Web Archiving*2006), 1-53.

Miller, C. & Shepherd, D.(2004) “Blogging as Social Action: A Genre Analysis of the Weblog.” Online article.
http://blog.lib.umn.edu/blogosphere/blogging_as_social_action_a_genre_analysis_of_the_weblog.html (accessed 16/08/2012)

NARA. *Significant Properties*. NARA, 2009.

Nardi, B. A., Schiano, D. J., Gumbrecht, M. and Swartz, L. Why we blog. *Communications of the ACM*, 47, 12 2004), 41-46.

O'Sullivan, C. Diaries, on-line diaries, and the future loss to archives; or, blogs and the blogging bloggers who blog them. *American Archivist*, 68, 1 2005), 53-73.

PARADIGM report [precise ref to be inserted]
http://www.paradigm.ac.uk/workbook/pdfs/08_digital_preservation.pdf (accessed 24/08/2012)

Pennock, M. and Davis, R. *ArchivePress: A really simple solution to archiving blog content*. CDL, City, 2009.

Ponniah, P. *Data modeling fundamentals: a practical guide for IT professionals*. Wiley-Blackwell, Hoboken, New Jersey, USA, 2007.

Pluempavarn, P. and Panteli, N. *Building social identity through blogging*. Palgrave Macmillan, City, 2008.

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

Rauch, C. Strodl, S. & Rauber, A. (2005). Deliverable 6.4.1: A framework for documenting the behaviour and functionality of digital objects and preservation strategies.
http://www.dpc.delos.info/private/output/DELOS_WP6_d641_final__vienna.pdf (accessed 28/09/2012).

Ross, S. and Gow, A. (1999) Digital archaeology? Rescuing Neglected or Damaged Data Resources. Bristol & London: British Library and Joint Information Systems Committee. ISBN 1-900508-51-6.

Ross, S (2000) “Changing Trains at Wigan: Digital Preservation and the Future of Scholarship.” London, UK: National Preservation Office (British Library). ISBN 0-7123-4717-8.

Ross, S. (2006) “Approaching Digital Preservation Holistically”, in A Tough and M Moss (eds.), Information Management and Preservation, (Oxford: Chandos Press)
Rothenberg, Jeff (1995). "Ensuring the Longevity of Digital Documents". *Scientific American* **272** (1).

Sacchi, S. and McDonough, J. P. *Significant properties of complex digital artifacts: open issues from a video game case study*. ACM, City, 2012.

Scape project (2012b) *D15.1 Web content executable workflows for experimental execution*. Project deliverable.
http://www.scape-project.eu/wp-content/uploads/2012/05/SCAPE_D15.1_ONB_v1.0.pdf

Scape project (2012a) *D12.1 Preservation Watch Component Architecture*. Project deliverable.
http://www.scape-project.eu/wp-content/uploads/2012/01/SCAPE_D12.1_TUW_V1.0.pdf

Sesink, L., R. van Horik, and H. Harmsen (2008), *Data Seal of Approval*. Data Archiving and Networked Services (DANS), Den Haag, The Netherlands. Available from: http://www.datasealofapproval.org/

Sheble, L., Choemprayong, S. and Hank, C. *Surveying bloggers' perspectives on digital preservation: Methodological issues*. City, 2007.

Siles, I. (2012) "Web Technologies of the Self: the Arising of the Blogger Identity." *Journal of Computer-Mediated Communication*, Volume 17 Issue 4 pages 408-421 July 2012.
http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2012.01581.x/abstract
(accessed 16/08/2012)

SPRUCE report [precise ref to be inserted]

Stepanyan, K., Joy, M., Cristea, A., Kim, Y., Pinsent, E. and Kopidaki, S. *D2.2 Report: BlogForever Data Model*. 2011.

Strodl, S., Rauch, C., Rauber, A., Hofman, H., Debole, F., Amato, G. (2006) "The DELOS Testbed for Choosing a Digital Preservation Strategy." *Proceedings of the 9th International Conference on Asian Digital Libraries*, Kyoto, Japan, November 27-30, 2006.

Thibodeau, K. (2002) "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years" *The State of Digital Preservation: An International Perspective, Conference Proceedings*, Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub107/thibodeau.html

Tyan Low, J. *A literature review: What exactly should we preserve? How scholars address this question and where is the gap.* University of Pittsburgh, Pennsylvania, USA., City, 2011.

Wf4Ever report [precise ref to be inserted]
http://repo.wf4ever-project.org/dlibra/docmetadata?id=28&from=pubstats (accessed 24/08/2012)

Wilkinson, D. & Thelwall, M. (2010) "Social network site changes over time: The case of MySpace." Journal of the American Society for Information Science and Technology, 61(11), 2311 -2323.
http://www.scit.wlv.ac.uk/~cm1993/papers/SNS_changes_over_time_Preprint.doc
(preprint accessed 16/08/2012)

Wilson, A. (2005) "A Performance Model and Process for Preserving Digital Records for Long-term Access" Archiving 2005, Volume 2, ISBN / ISSN: 0-89208-255-0, 20-25.

Wilson, A. *Significant Properties: Report*. 2007.

Yeo, G. 'Nothing is the same as something else': significant properties and notions of identity and originality. *Archival Science*, 10, 2 2010), 85-116.

A. Appendix A – Draft METS profile for BlogForever


```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<METS_Profile xmlns:xhtml="http://www.w3.org/1999/xhtml"
xsi:schemaLocation="http://www.loc.gov/METS_Profile/v2
http://www.loc.gov/standards/mets/profile_docs/mets.profile.v2-
0.xsd"
   xmlns="http://www.loc.gov/METS_Profile/v2"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    STATUS="provisional" REGISTRATION="unregistered">
<URI LOCTYPE="URL" ASSIGNEDBY="local"><!-- [HK:] I think that
the actual URL should appear in the attribute ID but I am not
sure. --
>http://www.blogforever.eu/standards/mets/profiles/BFArchivePro
fileV4-03062012.xml</URI>
<!-- the URI above does not exist. it is just an example. A
real URI representing the profile location within the
repository or at a public registry of profiles (such as that at
the library of congress) should be created. -->
<title>BlogForever Archive Mets Profile Version 0.3a</title>
<abstract>This profile is intended to be used to govern the
implementation of a Blog Archive repository deployed to meet
the BlogForever weblog preservation, management and
dissemination standards. The digital content governed by the
METS documents conforming to this profile may be of any type or
combination of types including, but not limited to: Blogs, Blog
Post, Blog Comment and Blog Page and associated linked content
comprising semi-structured text, documents, audio, video, and
images. This profile covers born-digital materials found within
weblogs intended for general reference use.</abstract>
<date>2012-05-03T12:00:00</date>
<contact ID="ct1">
<name></name>
<address></address>
<email></email>
</contact>
<contact ID="ct2">
<name></name>
<address></address>
<email></email>
</contact>
<related_profile>No related profile.</related_profile>
<profile_context><!-- [HK:] Should be filled. --
></profile_context>
<external_schema ID="ext01">
<name>MARCXML</name>
 <URL><!-- [HK:] I think that the actual URL should appear in
the attribute ID but I am not sure. --
>http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd</UR
L>
<context>Used for descriptive metadata (paths
"mets/dmdSec/mdWrap/xmlData").</context>
</external_schema>
<external_schema ID="ext02">
```

```
<name>PREMIS</name>
 <URL><!-- [HK:] I think that the actual URL should appear in
the attribute ID but I am not sure. --
>http://www.loc.gov/standards/premis/premis.xsd</URL>
<context>For preservation metadata, including preservation
level, actions, provenance, and rights. (paths
"mets/amdSec/sourceMD/mdWrap/xmlData",
"mets/amdSec/digiprovMD/mdWrap/xmlData" and/or
"mets/amdSec/rightsMD/mdWrap/xmlData").</context>
</external_schema>
<external_schema ID="ext03">
<name>textMD</name>
 <URL><!-- [HK:] I think that the actual URL should appear in
the attribute ID but I am not sure. --
>http://www.loc.gov/standards/textMD/index.html</URL>
<context>Technical metadata schema for text. (paths
"mets/amdSec/techMD/mdWrap/xmlData").</context>
</external_schema>
<external_schema ID="ext04">
<name>MIX</name>
 <URL><!-- [HK:] I think that the actual URL should appear in
the attribute ID but I am not sure. --
>http://www.loc.gov/standards/mix/</URL>
<context>Technical metadata schema for images. (paths
"mets/amdSec/techMD/mdWrap/xmlData").</context>
</external_schema>
<external_schema ID="ext05">
<name>AES57-2011</name>
 <URL><!-- [HK:] I think that the actual URL should appear in
the attribute ID but I am not sure. --
>http://www.aes.org/publications/standards/search.cfm?docID=84<
/URL>
<context>Technical metadata schema for audio. (paths
"mets/amdSec/techMD/mdWrap/xmlData").</context>
</external_schema>
<external_schema ID="ext06">
<name>MPEG-7-Version10</name>
<URL></URL>
<context>Technical metadata schema for video. (paths
"mets/amdSec/techMD/mdWrap/xmlData").</context>
</external_schema>
<external_schema ID="ext07">
 <name>documentMD</name>
 <URL><!-- [HK:] I think that the actual URL should appear in
the attribute ID but I am not sure. --
>http://www.fclaweb.fcla.edu/uploads/Lydia%20Motyka/FDA_documen
tation/documentMD.pdf</URL>
<context>Technical metadata schema for formats intended primary
for office documentations (e.g. presentations, word processing
software documents, spreadsheets). Associated file extensions
might include doc, ppt, docx, odt, xls, ods, odp, PDF).
Metadata standard developed by Chou and Goethals (2009),
adopted by Florida Digital Library and Harvard University
Library (paths "mets/amdSec/techMD/mdWrap/xmlData").</context>
</external_schema>
<external_schema ID="ext08">
```

```
<name>Preservation Metadata for Digital Collections</name>
 <URL><!-- [HK:] I think that the actual URL should appear in
the attribute ID but I am not sure. --
>http://www.nla.gov.au/preserve/pmeta.html</URL>
<context>Possible technical metadata schema for scripts.
Associated file extensions include js, rsd, rss, rdf. These
might be handled with structured or semi-structured text
instead. (paths "mets/amdSec/techMD/mdWrap/xmlData").</context>
</external_schema>
<description_rules>
 <!-- In addition to anything included here, see
structural_requirements within this profile. -->
 <!-- [HK:] The description_rules element is not allowed to
contain any requirement elements. -->
 <head ID="dmdSecDR"></head><!--  RELATEDMAT="ext01" -->
 <p xmlns="http://www.w3.org/1999/xhtml">The descriptive
metadata for the purpose of browsing, search, seeking and
discovery must be included in the "dmdSec" section of the METS
object and expressed using MARCXML (extension schema "ext01").
Additional schemas may be used but these must be explicitly
described in the "extension_schema" section. Any alternative
schema must be used within a separate "dmdSec" unless it is
being used to replace MARCXML system wide. In the latter case,
this new rule must be stated in the description rules.</p>
<!--
 <requirement ID="dmdSecDR"  RELATEDMAT="ext01">
The descriptive metadata for the purpose of browsing, search,
seeking and discovery must be included in the "dmdSec" section
of the METS object and expressed using MARCXML (extension
schema "ext01"). Additional schemas may be used but these must
be explicitly described in the "extension_schema" section. Any
alternative schema must be used within a separate "dmdSec"
unless it is being used to replace MARCXML system wide. In the
latter case, this new rule must be stated in the description
rules.
<requirement ID="recordTypeDR" RELATEDMAT="vc1">
The type of the record must be indicated in the MARCXML
embedded within the section "dmdSec" using the vocabulary
"vc1".
<requirement ID="recordTypeBlogDR">
A blog in the BlogForever archive is considered to be the
publication venue that provides a location for a collection of
intellectual entities including but not necessarily limited to
page, blog post, comment and embedded content. As such, the
blog record description will not include the list of blog
posts, pages, comments and categorised content. The association
between the blog and these items must be indicated as part of
the description metadata of each individual entity published
within the blog.
</requirement>
<requirement ID="recordTypeBlogPostDR"
RELATEDMAT="recordTypeBlogDR">
The description of each blog post ingested into the repository
must include a link to the parent blog.
</requirement>
```

```
<requirement ID="recordTypePageDR"
RELATEDMAT="recordTypeBlogDR">
The description of each page ingested into the repository must
include a link to the parent blog.
</requirement>
<requirement ID="recordTypeCatContentDR"
RELATEDMAT="recordTypeBlogDR">
The description of each embedded content ingested into the
repository must include a link to the parent blog, comment,
post or page.
</requirement>
<requirement ID="recordTypeCommentDR"
RELATEDMAT="recordTypeBlogDR">
The description of each comment ingested into the repository
must include a link to the parent blog post, comment, and/or
page.
</requirement>
</requirement>
<requirement ID="BlogTypeDR" RELATEDMAT="vc2">
The type of the blog and its entries must be indicated in the
MARCXML embedded within the section "dmdSec" using the
vocabulary "vc2".
</requirement>
<requirement ID="BlogTopicDR" RELATEDMAT="vc3">
The topic or subject area of the blog and its entries must be
indicated in the MARCXML embedded within the section "dmdSec"
using the vocabulary "vc3".
</requirement>
<requirement ID="BlogStatusDR" RELATEDMAT="vc4">
The HTTP response state of the blog and its entries must be
indicated in the MARCXML embedded within the section "dmdSec"
using the vocabulary "vc4".
</requirement>
</requirement>

<requirement ID="techMDDR1" RELATEDMAT="ext02">
The highest level technical metadata must be expressed using
PREMIS (extension schema "ext02") unless it has been agreed
that an alternative schema replace PREMIS. In the latter case,
the new schema should be listed in the "extension schema"
section, and a new rule stated to this effect in the
description rules. If it is decided that more than one schema
be used at any one time, the rules by which this must be
implemented must be included in the "structural_requirements"
section.
<requirement ID="techMDDR2" RELATEDMAT="vc11">
Technical metadata related to MIME media types (controlled
vocabulary vc11) must be included wherever possible.
<requirement ID="techMDDR2" RELATEDMAT="ext02">
Technical metadata related to MIME media types (controlled
vocabulary vc11) must be included wherever possible.
Any such metadata must be included in the "techMD" section of
the METS object using the "objectCharacteristicsExtension"
element of PREMIS (extension schema "ext02").
</requirement>
</requirement>
```

```
<requirement ID="techMDDR3" RELATEDMAT="ext03">
It is recommended that technical metadata specific to "text"
media type use textMD (extension schema "ext03").
</requirement>
<requirement ID="techMDDR4" RELATEDMAT="ext04">
It is recommended that technical metadata specific to "image"
media type use MIX (extension schema "ext04").
</requirement>
<requirement ID="techMDDR5" RELATEDMAT="ext05">
It is recommended that technical metadata specific to "audio"
media type use AES (extension schema "ext05").
</requirement>
<requirement ID="techMDDR4" RELATEDMAT="ext06">
It is recommended that technical metadata specific to "video"
media type use MPEG-7 (extension schema "ext06").
</requirement>
<requirement ID="techMDDR4" RELATEDMAT="ext07">
It is recommended that technical metadata specific to
"application" use docMD where it is possible (extension schema
"ext07").
</requirement>
</requirement>

<requirement ID="dateTimeDR" RELATEDMAT="vc9">
The value of any instance of a date and/or time element within
this profile and any sections of associated METS objects must
be expressed using the controlled vocabulary "vc9".
</requirement>
<requirement ID="languageDR" RELATEDMAT="vc5">
The value of any instance of a language field element within
this profile and any sections of associated METS objects must
be expressed using the controlled vocabulary "vc5".
</requirement>
<requirement ID="encodingDR" RELATEDMAT="vc6">
The value of any field referring to language encoding or
character set within this profile and any sections of
associated METS objects must be expressed using the controlled
vocabulary "vc6".
</requirement>
<requirement ID="countryDR" RELATEDMAT="vc7">
The value of any field refering to country location within this
profile and any sections of associated METS objects must be
expressed using the controlled vocabulary "vc7".
</requirement>

<requirement ID="rightsDR1" RELATEDMAT="ext02">
The rights associated to a digital object must be expressed
using the PREMIS schema (extension schema "ext02"). This must
be embedded within the "rightsMD" section of the METS object.
The link between the object and rightsMD will be expressed as
part of the technical metadata of the object.
<requirement ID="rightsDR1" RELATEDMAT="vc8">
The value of fields relating to rights metadata must use the
controlled vocabulary specified "vc8".
</requirement>
</requirement>
```

```
<requirement ID="eventDR1" RELATEDMAT="ext02">
The description of repository events must use the PREMIS
(extension schema "ext02") event description standards.
<requirement ID="eventDR2" RELATEDMAT="vc10">
The values associated to the schema used in describing
repository events must use the controlled vocabulary "vc10".
</requirement>
</requirement>

-->
</description_rules>

<controlled_vocabularies>

<vocabulary ID="vc1">
<name>Record type standard</name>
 <!-- [HK:] There should also be provided an agency responsible
for maintaining the vocabulary (according to the METS
description).-->
 <URI><!-- [HK:] A URI should be indicated in the ID attribute.
--></URI>
 <description><xhtml:p>This vocabulary is intended for
indicating record type where records are varying in type (e.g.
blog, blog entry - and subtypes of the
entries).</xhtml:p></description>
<!--
 <comment>If a standard does not exist, we should define one
and create a URI where the standard is expressed formally. An
example is provided below:
<example>
<recordType>
<p>blog</p>
<p>entry
<p>blog post</p>
<p>comment</p>
<p>home page</p>
<p>about page</p>
<p>unclassified page</p>
</p>
</recordType>
</example>
</comment>
-->
</vocabulary>

<vocabulary ID="vc2">
<name>Blogtype Taxonomy</name>
<URI></URI>
 <description><xhtml:p>This is for expressing blog
type.</xhtml:p></description>
<!-- <comment>What blog type means has to be clarified but the
field was mentioned in the data model. We should clarify what
it means, create a taxonomy. On a simplified level this could
indicate whether the blog is a corporate blog or a personal
blog (or research, government and/or general interest). How
this can be extracted is unclear. While it is possible to use
```

```
the URL to determine where it is hosted, it may not be possible
to authoritatively say to which category it belongs. If we
nevertheless decide to use this and there is no such schema
already in use, a URI and description should be created for it
on the archive site.</comment> -->
</vocabulary>

<vocabulary ID="vc3">
<name>Blog Topic Taxonomy</name>
<!-- <comment>Can we use LC subject headings? If not what
alternative is available? LCSH URI is provided below.</comment>
-->
<URI>http://id.loc.gov/authorities/subjects.html</URI>
 <description><xhtml:p>This is for high level content
description, e.g. general subject area of the blog, blog post,
comment or page.</xhtml:p></description>
<!-- <comment>LCSH may not be appropriate for describing web
content, If it is deemed not adequate and there is no
alternative schema already in use, a URI and description should
be created for it on the archive site.</comment> -->
</vocabulary>

<vocabulary ID="vc4">
<name>Some standard expressing blog source web URI status (e.g.
http response code).</name>
<URI>http://www.w3.org/Protocols/HTTP/HTRESP.html</URI>
 <description><xhtml:p>This is to give some indication of
whether the blog might still be active as part of the web and
to pre-empt problems that might be occur in its continued
preservation.</xhtml:p></description>
</vocabulary>

<vocabulary ID="vc5">
<name>ISO 639-2</name>
<URI>http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogu
e_detail.htm?csnumber=22109</URI>
 <description><xhtml:p>Standard code for expressing
languages.</xhtml:p></description>
</vocabulary>

<vocabulary ID="vc6">
<name>ISO 8859</name>
<URI>http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogu
e_detail.htm?csnumber=28263</URI>
 <description><xhtml:p>ISO standard for describing
charactersets. ISO 8859-1 to ISO 8859-
16.</xhtml:p></description>
<!-- <comment>Alternatively, the IANA encoding vocabulary could
be used. See http://www.iana.org/assignments/character-sets.
Whichever schema is used, it must be agreed upon and specified
here.</comment> -->
</vocabulary>

<vocabulary ID="vc7">
<name>ISO 3166-1</name>
```

```
<URI>http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogu
e_detail.htm?csnumber=39719</URI>
 <description><xhtml:p>This is a standard code for expressing
country location.</xhtml:p></description>
</vocabulary>

<vocabulary ID="vc8">
<name>Rights and Licenses Vocabulary</name>
<URI></URI>
<description> </description>
<!-- <comment>
If there is no standard that we can draw upon one should be
agreed upon and published at a URI within the repository.
The following example is constructed from the PREMIS
recommendation:
<example>
<rightsBasis>
<p>copyright</p>
<p>license</p>
<p>statute</p>
</rightsBasis>
<rightsGranted>
<act>
<p>replicate = make an exact copy</p>
<p>migrate = make a copy in a different file format</p>
<p>modify = make a version different in content</p>
<p>use = read without copying or modifying</p>
<p>disseminate = copy for use outside the repository</p>
<p>delete = remove from the repository</p>
</act>
</rightsGranted>
<copyrightInformation>
<copyrightStatus>
<p>copyrighted = Under copyright.</p>
<p>publicdomain = In the public domain.</p>
<p>unknown = Copyright status of the resource is unknown.</p>
</copyrightStatus>
<copyrightJurisdiction>
ISO-3166 country codes. Same vocabulary for other fields
expressing country codes within every METS object.
</copyrightJurisdiction>
<copyrightDeterminationDate>
ISO-8601 for dates, GMT, and GMT offset. Same vocabulary for
other dates and times appearing within every METS object.
</copyrightDeterminationDate>
</copyrightInformation>
</example>
</comment>
-->
</vocabulary>

<vocabulary ID="vc9">
  <name>dateTime</name>
  <maintenance_agency>W3C</maintenance_agency>
  <URI>http://www.w3.org/TR/xmlschema-2/#dateTime</URI>
```

```
  <context>The vocabulary should be used in a METS object in
these elements where timestamps have to be expressed. For more
information, see the XML Schema specification for METS objects:
http://www.loc.gov/standards/mets/mets.xsd</context>
  <description>
    <xhtml:p>The vocabulary is used to express timestamps. It
consists of values for date and time. Time zone can be added as
well.</xhtml:p>
  </description>
 </vocabulary>

<vocabulary ID="vc10">
<name>Repository Events</name>
<URI></URI>
 <description><xhtml:p>A standard for describing management,
procedural or preservation events that occur within the
repository.</xhtml:p></description>
<!-- <comment>
<p>Suggested initial list from PREMIS:
<example>
<eventType>
<p>capture = the process whereby a repository actively obtains
an object</p>
<p>compression = the process of coding data to save storage
space or transmission time</p>
<p>creation = the act of creating a new object</p>
<p>deaccession = the process of removing an object from the
inventory of a repository</p>
<p>decompression = the process of reversing the effects of
compression</p>
<p>decryption = the process of converting encrypted data to
plaintext</p>
<p>deletion = the process of removing an object from repository
storage</p>
<p>digital_signature_validation = the process of determining
that a decrypted digital signature matches an expected
value</p>
<p>dissemination = the process of retrieving an object from
repository storage and making it available to users</p>
<p>fixity_check = the process of verifying that an object has
not been changed in a given period</p>
<p>ingestion = the process of adding objects to a preservation
repository</p>
<p>message_digest_calculation = the process by which a message
digest (â€œhashâ€) is created</p>
<p>migration = a transformation of an object creating a version
in a more contemporary format</p>
<p>normalization = a transformation of an object creating a
version more conducive to preservation</p>
<p>replication = the process of creating a copy of an object
that is, bit-wise, identical to the original</p>
<p>validation = the process of comparing an object with a
standard and noting compliance or exceptions</p>
<p>virus_check = the process of scanning a file for malicious
programs</p>
</eventType>
```

```
</example>
</p>
<p>
New schema suggested below based on the recommendations in
PREMIS version 2.1 and preservation strategies described in the
tutorial developed by Cornell University
(http://www.dpworkshop.org/dpm-eng/terminology/strategies.html)
and discussion within the PARADIGM project
(http://www.paradigm.ac.uk/workbook/metadata/preservation-
event.html):
<example>
<entityType>
<p>replication = the process of making an exact copy
 <p>bitstream copying</p>
 <p>LOCKSS</p>
 <p>transactional replication</p>
</p>
<p>refreshing = moving content to a new medium because there is
danger/evidence of deterioration in the existing medium or
because another medium is deemed more durable. Analog backup
could be considered to be a form of this</p>
<p>modification = the process of altering the content or format
of the object
 <p>migration = make a copy in a different file format</p>
 <p>content_alteration = make a version different in content of
the information object</p>
 <p>metadata_modification = modifying a metadata object</p>
 <p>deaccession = the process of removing an object from the
inventory of a repository</p>
 <p>compression = the process of coding data to save storage
space or
transmission time</p>
 <p>decompression = the process of reversing the effects of
compression</p>
 <p>decryption = the process of converting encrypted data to
plaintext</p>
 <p>normalization = a transformation of an object creating a
version more conducive to preservation</p>
 <p>restoration = recovering content from backup or by other
means when corruption is discovered</p>
</p>
<p>reading = the act of viewing or processing without copying
or modifying
 <p>information_processing = algorithmic analysis of
information to extract, synthesise or create content</p>
 <p>rendering = displaying information on a hardware device
using a software, and/or emulator</p>
 <p>message_digest_calculation = the process by which a message
digest (â€œhashâ€) is created</p>
 <p>checking = verifying standards, values and state to
determine object conformance to a target standard, value,
and/or state
  <p>virus_check = the process of scanning a file for malicious
programs</p>
  <p>format_validation = the process of comparing an object
with a format standard to assess compliance</p>
```

```
  <p>fixity_check = the process of verifying that an object
hash value has not been changed in a given period</p>
  <p>digital_signature_validation = the process of determining
that a decrypted digital signature matches an expected
value</p>
 </p>
</p>
<p>dissemination = the process of distributing information for
use outside the repository</p>
<p>deletion = the process of removing from the repository</p>
<p>creation = the process of creating an object
 <p>content_creation = creation of an information object</p>
 <p>metadata_creation = creation of a metadata object</p>
 <p>canonicalisation = creation of a profile of an object that
can be used to assess whether the essential characteristics of
the object remains intact</p>
 <p>implement_emulator = enable the reproduction of the
essential characteristics and performance of a computing
environment</p>
 <p>retarget_code = translate code on one environment to work
in an another environment</p>
 <p>deploy_self_aware_object</p>
</p>
<p>capture = the process whereby a repository actively obtains
an object</p>
<p>ingestion = the process of adding objects to a preservation
repository</p>
<p>annotation = the process of associating meta-information
regarding an object
 <p>encapsulation = grouping an object and its associated
metadata into a single object</p>
 <p>contextualisation = making explicit a relationship between
an object and other information such as its associated metadata
or canonicalisation</p>
</p>
</entityType>
</example>
</p>
</comment>
-->
</vocabulary>

<vocabulary ID="vc11">
<name>MIME media type</name>
<URI>http://www.iana.org/assignments/media-
types/index.html</URI>
 <description><xhtml:p>IANA list of mime media types. See also
http://www.ietf.org/rfc/rfc2046.txt?number=2046</xhtml:p></desc
ription>
<!-- <comment>This is used for each file within a record being
described as PREMIS format description included in the child
element "techMD" of the METS "amdSec" section. This should
guide the selection of any objectCharacteristicsExtension of
the PREMIS schema describing specific format
characteristics.</comment> -->
</vocabulary>
```

```
</controlled_vocabularies>


<structural_requirements>
 <!--
<requirement ID="coverage">
This profile outlines the requirements in creating a record for
a blog as a collection of associated pages, and/or, for each
individual page within the blog.
</requirement>
 -->
 <metsRootElement>
   <!-- Every METS object in the repository must contain a root
"mets" element. -->
  <requirement ID="rootID">
   <description><xhtml:p>Every root "mets" element must contain
an attribute "ID" whose value represents a unique descriptive
meaningful identifier in the repository for the corresponding
METS object. This identifier will commonly consist of a
repository URI and a meaningful human-readable descriptive
local URI within the repository for the METS
object.</xhtml:p></description>
  </requirement>
  <requirement ID="rootOBJID">
   <description><xhtml:p>Every root "mets" element must contain
an attribute "OBJID" whose value represents a globally unique
identifier of the object within the repository that the current
METS object is recording. This identifier will commonly consist
of a repository URI and a meaningful local URI within the
repository indicating the location of the object. The URI must
remain the same with respect to all repository functions
including the publication of LinkedData or other ontological
representations. If the URI should be changed, the old IDs must
be retained using the "altRecordID" attribute of the metsHdr
element.</xhtml:p></description>
  </requirement>
  <requirement ID="rootLABEL">
   <description><xhtml:p>Every root "mets" element must contain
an attribute "LABEL" whose value represents a descriptive
human-readable name for the object that the Mets object is
recording. For example, for a Blog object this could be the
title of the source weblog.</xhtml:p></description>
  </requirement>
  <requirement ID="rootPROFILE">
   <description><xhtml:p>Every root "mets" element must contain
the attribute "PROFILE" whose value represents the URI of this
profile, i.e. the profile that specifies the rules and schemas
and vocabularies with which the METS object was
created.</xhtml:p></description>
  </requirement>
  <requirement ID="rootSCHEMA">
   <description><xhtml:p>The root "mets" elements must include
locations for the METS object schema, extension schemas being
used in the METS object.</xhtml:p></description>
  </requirement>
 </metsRootElement>
```

```
<metsHdr>
 <!--
 <requirement ID="header">
  Every METS object in the repository must contain a header
"metsHdr" element.
 -->
  <requirement ID="metsHdrCREATEDATE">
   <description><xhtml:p>Every header element "metsHdr" must
contain an attribute "CRATEDATE" representing the date and time
that the METS object was first created. The value must follow
the agreed vocabularies for expressing date and time
(vocabulary ID VC9 and VC10)</xhtml:p></description>
  </requirement>
  <requirement ID="metsHdrLASTMODDATE">
   <description><xhtml:p>Every header element "metsHdr" must
contain an attribute "LASTMODDATE" representing the date and
time that the METS object was last modified. The value must
follow the agreed vocabularies for expressing date and time
(vocabulary ID VC9 and VC10).</xhtml:p></description>
  </requirement>
  <!-- <requirement ID="metsHdrID"> -->
   <requirement ID="metsHdrRepeatID">
    <description><xhtml:p>It is recommended that the "ID" and
"OBJID" of the root "mets" element attribute be repeated as
metHdr attributes.</xhtml:p></description>
   </requirement>
   <requirement ID="metsHdrAltRecordID">
    <description><xhtml:p>Should there be a change in the OBJID
of METS object for some unavoidable reason, the old object ID
must be retained as a child element "altRecordID" of the
"metsHdr" element. An attribute "TYPE" must be used to indicate
the type of the old record ID (e.g.
DOI).</xhtml:p></description>
   </requirement>
  <!-- </requirement> -->
  <requirement ID="metsHdrAgent">
   <description><xhtml:p>Any agents responsible for the
modification of the METS object should ideally be indicated
within the child element "agent" of the header element
"metsHdr". At least one agent should be specified as the
custodian of the METS object.</xhtml:p></description>
  </requirement>
  <requirement ID="metsHdrAgentID">
   <description><xhtml:p>Any metsHdr agent must have an
attribute "ID" whose value represents a global URI for the
agent whether this is a software, service, organisation, or
person.</xhtml:p></description>
  </requirement>
  <requirement ID="metsHdrAgentROLE">
   <description><xhtml:p>Any metsHdr agent must have an
attribute "ROLE" whose value represents what role the agent
played in relation to the METS object. Every metsHdr must
contain at least one agent whose "ROLE" attribute value is
"CUSTODIAN".</xhtml:p></description>
  </requirement>
```

```
<!-- </requirement> -->
</metsHdr>

<dmdSec>
 <requirement ID ="dmdSec">
  <description><xhtml:p>Every METS object in the repository
must contain at least one descriptive metadata section element
"dmdSec". While there might be more than one "dmdSec", only one
PRIMARY description should be given.</xhtml:p></description>
 </requirement>
  <requirement ID="dmdSecID">
   <description><xhtml:p>Every instance of "dmdSec" must be
used with an attribute "ID" whose value represents a unique ID
for the presented descriptive metadata.</xhtml:p></description>
  </requirement>
  <requirement ID="dmdSecWrap">
   <description><xhtml:p>Every instance of a single metadata
type should be wrapped in the element
"mdWrap".</xhtml:p></description>
  </requirement>
   <requirement ID="dmdSecWrapMDTYPE">
    <description><xhtml:p>Every instance of "mdWrap" must be
used with an attribute "MDTYPE" whose value represents the
metadata schema (in its abbreviated form) that is being used to
express the contents encapsulated by
"mdWrapper".</xhtml:p></description>
   </requirement>
   <requirement ID="dmdSecWrapMIMETYPE">
    <description><xhtml:p>Every instance of "mdWrapper" must be
used with an attribute "MDTYPE" whose value represents a single
metadata schema that is being used in the "mdWrapper"
section.</xhtml:p></description>
   </requirement>
   <requirement ID="dmdSecWrapDATATYPE">
    <description><xhtml:p>The contents encapsulated by the
element "mdWrapper" must be either expressed in XML, using the
child element "xmlData", or in binary form, using the child
element "binData".</xhtml:p></description>
   </requirement>
</dmdSec>

<amdSec>
 <requirement ID="amdSec">
  <description><xhtml:p>Every METS object must be associated to
at least one administrative metadata section element "amdSec".
Each "amdSec" section must contain at least one "techMD", one
"rightsMD", and one "digiprovMD"
section.</xhtml:p></description>
 </requirement>
 <requirement ID="amdSecTechMD">
  <description><xhtml:p>Every METS object contains at least one
"techMD" element as a child element of "amdSec", specifying the
technical aspects of a file within the object associated to the
METS object.</xhtml:p></description>
 </requirement>
 <requirement ID="amdSecTechMDID">
```

```
  <description><xhtml:p>Every METS "techMD" element must come
with an attribute "ID" that identifies it uniquely within the
METS object. This ID will be used to link the object to the
technical metadata.</xhtml:p></description>
 </requirement>
 <requirement ID="amdSecTechMDOBject">
  <description><xhtml:p>The TechMD is where all the object
information relevant to management and technical processes will
be specified. A new techMD section will be created to
correspond to each file within the object associated to the
METS object. For example, any image described as part of a blog
post with have its own corresponding techMD section. This
metadata (object identifier, preservation level, object
characteristics, storage information, environment information,
signature information, relationships to other objects - not
hyper reference, links to events - described within
"digiProvMD" section, and links to rights -described within
"rightsMD" section) must be wrapped in the PREMIS schema and
further format characteristics should be included wherever
possible using extension schemas listed in this profile and
these must be wrapped in "objectCharacteristicsExtension" of
the PREMIS metadata section.</xhtml:p></description>
 </requirement>
 <requirement ID="amdSecRightsMD">
  <description><xhtml:p>Every METS object contains at least one
"rightsMD" element as a child element of "amdSec", specifying
the legal mandates associated to the digital object described
by the METS object.</xhtml:p></description>
 </requirement>
 <requirement ID="amdSecRightsMDID">
  <description><xhtml:p>Each "rightsMD" must be assigned with
an "ID" attribute assigning a unique ID for the "rightsMD"
section within the corresponding METS
object.</xhtml:p></description>
 </requirement>
 <requirement ID="amdSecRightsMDExpression">
  <description><xhtml:p>Each "rightsMD" must be expressed using
the PREMIS rights description and rights expression vocabulary
vc8.</xhtml:p></description>
 </requirement>
 <requirement ID="amdSecDigiProvMD">
  <description><xhtml:p>Every METS object contains at least one
"digiprovMD" element as a child element of "amdSec", specifying
the provenance of the object associated to the METS
object.</xhtml:p></description>
 </requirement>
 <requirement ID="amdSecDigiProvMDID">
  <description><xhtml:p>Every "digiprovMD" must be assigned an
attribute "ID" whose value is a unique identifier of the
metadata section within the METS
object.</xhtml:p></description>
 </requirement>
 <requirement ID="digiProvMDEvent">
  <description><xhtml:p>Each repository event must be recorded
here using he language of a PREMIS event and assigned an
```

```
attribute "ID" whose value functions as a unique within the
repository.</xhtml:p></description>
 </requirement>
</amdSec>

<fileSec>
 <requirement ID="fileSec">
  <description><xhtml:p>Every METS object in the repository
must contain at least one "fileSec" element and associated
child element "fileGrp", listing files that are grouped
together to form a single representation of an intellectual
entity (e.g. image files of each page of a book that come
together to comprise the single intellectual book entity).
There can be several "fileGrp" elements associated to several
representations of the same intellectual entity (e.g. a scanned
image and a word document representing the same
letter).</xhtml:p></description>
 </requirement>
  <requirement ID="fileMD">
   <description><xhtml:p>Every file object encapsulated by
"fileSec" must be associated to one and only one of the
"techMD" section.</xhtml:p></description>
  </requirement>
</fileSec>

<structMap>
 <requirement ID="structMap">
  <description><xhtml:p>Every METS object in the repository
must contain at least one "structMap" element specifying how
the files described in the child element "fileGrp" of the
element "fileSec" are organised in relation to each other (e.g.
the order of the book pages each of which are represented as an
image file).</xhtml:p></description>
 </requirement>
  <requirement ID="structMapDIV">
   <description><xhtml:p>The divisions encapsulated by
"structMap" is expressed in an explicit hierarchical structure
using the "div" element.</xhtml:p></description>
  </requirement>
    <requirement ID="structMAPDIVResources">
     <description><xhtml:p>Resources linked within each "div"
section such as video, audio, image, other files (e.g. pdf
files), scripts, libraries, databases and links to other
webpages must be declared as "DEPENDENT_WEB_RESOURCE" using the
"TYPE" attribute of the "div" element.</xhtml:p></description>
    </requirement>
</structMap>

<structLink>
 <requirement ID="structLink">
  <description><xhtml:p>All hyperlink references from one Blog
Post to another Blog Post within the repository must be exposed
within the "structLink" section of the METS object using the
childe element "smLink" and its attributes "xlink:from" and
"xlink:to".</xhtml:p></description>
 </requirement>
```

```
 <requirement ID="behaviour">
  <description><xhtml:p>While it is not a requirement to
include a "behaviour" section within the METS object, it is
recommended that behaviour is described wherever possible. For
example, expected behaviour for clicking on a link is
recommended to be included. The effects of JavaScripts are also
to be considered for inclusion.</xhtml:p></description>
 </requirement>
</structLink>

</structural_requirements>

 <technical_requirements>
</technical_requirements>

 <!-- <tools> -->
 <tool ID="tl1">
  <name>JHOVE</name>
  <URI>http://hul.harvard.edu/jhove/</URI>
  <description><xhtml:p>JOHVE can be used for extracting
technical metadata from embedded content. The standard
representation information reported by JHOVE includes: file
pathname or URI, last modification date, byte size, format,
format version, MIME type, format profiles, and optionally,
CRC32, MD5, and SHA-1 checksums (information resp. at
http://hul.harvard.edu/jhove/references.html#crc32,
http://hul.harvard.edu/jhove/references.html#md5,
http://hul.harvard.edu/jhove/references.html#sha1). The initial
release of JHOVE includes modules for arbitrary byte streams,
ASCII and UTF-8 encoded text, GIF, JPEG2000, and JPEG, and TIFF
images, AIFF and WAVE audio, PDF, HTML, and XML; and text and
XML output handlers.</xhtml:p></description>
 </tool>
 <!-- </tools> -->

 <Appendix NUMBER="0"><xhtml:br/><!-- A profile must contain an
appendix containing an example METS document which conforms to
the requirements set out in the profile. Profile authors should
note that in order to insure that the completed profile
document is valid, any namespace and schemaLocation
declarations contained in the root <mets> element should be
moved to the root <METS_Profile> element. --></Appendix>

</METS_Profile>
```

B. Appendix B – Example Blog Post in METS

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- <Appendix NUMBER="1" LABEL="Simple METS example for a
Blog"> -->
<!-- The following values in the mets root element is not
really existing values yet. this is just to illustrate METS.
The ID must point to where the mets object will reside in the
repository. The objid must point to the location of the actual
blog in the repository. the profile must point to where the
METS profile resides - not the mets object associated to the
blog but the profile that tells you the rules according to
which each blog is described in the repository. the METS schema
location and any other general schemas must be indicated here
but specific schemas (e.g. PREMIS) must be indicated in the
corresponding sections. -->

<!--
    Important changes in the METS part:
    1.) An ID must not contain ":", "/", "=", "-", "?", or
"&amp;"
    2.) rights sections have to occur before digiprov sections
and not after
    3.) mdWrap MDTYPE="MARCXML" changed to MDTYPE="MARC"
    4.) FLocat LOCTYPE="URI" changed to LOCTYPE="URL"
    5.) It is not allowed to use the same ID twice
    6.) In a "file" element the value of a ADMID attribute has
to be a existing ID in the document (because it references to
an ID). Therefore, in this document ADMID="techMD2" was not
valid for a file. I changed it to ADMID="post_snapshot_master"
because this is the ID of the second techMD in this document.
Should it reference to this section?
    7.) In a "area" element the value of a FILEID attribute has
to be the ID of an existing file element in the document
(because it references to an ID of file element). Therefore,
"FID1", "FID2", "FID3", etc. were not valid values for FILEID
in the area element in this document. I changed it to existing
IDs. Should be reviewed if it links to the correct files.

    Important changes in PREMIS part:
    1.) object is abstract and needs further be defined with
the attribute xsi:type. For example: If the PREMIS type of the
object is "file", the object element needs the attribute
xsi:type="file"
    2.) objectCategory is not used in object. The category is
indicated instead by the type attribute (see 1.).
    3.) The element linkingEventIdentifier must have the
subelements linkingEventIdentifierType and
linkingEventIdentifierValue. It is not allowed that the
linkingEventIdentifier has characters as child. Therefore, the
values of linkingEventIdentifier were put in a subelement
linkingEventIdentifierValue.
    4.) Same problem as in 3.) for the element
linkingRightsStatementIdentifier.
```

```
    5.) A premis element needs a version attribute. Therefore,
the attribute version="2.2" has been added to the premis
elements.
    6.) A premis element must have an object element as the
first subelement. It is not allowed that a premis element
consists of only an agent, event, or rights element. Therefore,
placeholder object elements has been added.
    7.) The element copyrightDeterminationDate has been renamed
to copyrightStatusDeterminationDate
    8.) The element eventDateTime must contain a value. An
empty eventDateTime element is not allowed. Therefore, the
value 0001-01-01T00:00:00 has been added as a placeholder.
    9.) The element eventOutcomeInformation must not be empty.
Therefore, empty eventOutcomeInformation elements has been
commented out.

    Remarks for MARC part:
    1.) Intended URI for the repository, date_captured,
previous_version, next_version, versions does not exist as a
datafield
    2.) Why is the status_code distinguished for the blog and
the crawler?
    3.) Parent blog name and parent blog should probably appear
in a separate record for the blog but not in the same record as
the blog post.
    4.) The example indicated that previous_URI and next_URI
should contain URIs in the repository. However, is it not more
reasonable to include the original URIs?
    5.) What are possible subfield codes for field 532
(charset/encoding)?

 -->

<mets
    ID="http___blogforever.eu_mets_gowers"
    OBJID="http://blogforever.eu/blog/gowers"
    LABEL="Gowers's Weblog"

PROFILE="http://blogforever.eu/metsprofile/BFArchiveProfileV030
62012.xml"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns="http://www.loc.gov/METS/"
    xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/mets.xsd
    info:lc/xmlns/premis-v2
http://www.loc.gov/standards/premis/premis.xsd">

<metsHdr ID="BF_Blog_EXAMPLE_1" CREATEDATE="2012-05-
02T14:43:02" LASTMODDATE="2012-05-03T11:36:00">
<agent ROLE="CUSTODIAN" TYPE="ORGANIZATION">
<name>BlogForever Consortium</name>
</agent>
<altRecordID TYPE="URI"></altRecordID>
</metsHdr>
```

```
<dmdSec ID="dmdMD2" STATUS="PRIMARY_DMDSEC" CREATED="2012-06-
06T14:43:00-06:00" ADMID="digiProvMD6">
<mdWrap MDTYPE="MARC">
<xmlData>
    <record xmlns="http://www.loc.gov/MARC21/slim"
xsi:schemaLocation="http://www.loc.gov/MARC21/slim
http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd">
        <leader><!-- the following number is just a fictive
example -->00001AAAAA2200001AAA4500</leader>
        <datafield tag="245" ind1=" " ind2=" ">
            <subfield code="a">EPSRC update update</subfield>
            <subfield code="b">Not available</subfield>
        </datafield>
        <datafield tag="100" ind1=" " ind2=" ">
            <subfield code="a">Tim Gowers</subfield>
        </datafield>
        <datafield tag="520" ind1=" " ind2=" ">
            <subfield
code="u">http://gowers.wordpress.com/2012/05/31/epsrc-update-
update/<!-- The source URI from where it was captured. --
></subfield>
        </datafield>

<!-- <URI>
http://blogforever.eu/blogs/gowers/2012/05/31/epsrc-update-
update/ -->
<!-- This is intended to be the URI in the repository -->
<!-- </URI> -->

        <datafield tag="269" ind1=" " ind2=" ">
            <subfield code="c">2012-05-31T00:00:00</subfield>
        </datafield>
        <datafield tag="260" ind1=" " ind2=" ">
            <subfield code="m">Not available</subfield>
        </datafield>
<!-- <date_captured>
08062012UTC14:00-06:00
</date_captured> -->
<!-- <previous_version>
Not available -->
<!-- if it exists -->
<!-- </previous_version> -->
<!-- <next_version> -->
<!-- if it exists -->
<!-- </next_version> -->
<!-- <versions>
http://blogforever.eu/gowers/2012/05/31/epsrc-update-
update/versions -->
<!-- uri for a page where all versions of this post is
available -->
<!-- </versions> -->

        <datafield tag="952" ind1=" " ind2=" ">
            <subfield code="a">200</subfield>
        </datafield>
<!-- <status_code>
```

```
<browserRequestResponse>
200 -->
<!--  HTTP request response code (requested as browser). The
above code is fictitious. -->
<!-- </browserRequestResponse> -->
<!-- <crawlerRequestResponse>
301 -->
<!--  HTTP request response code (requested as crawler). The
above code is fictitious. -->
<!-- </crawlerRequestResponse>
</status_code> -->

        <datafield tag="788" ind1=" " ind2=" ">
            <subfield code="b">0</subfield>
            <subfield code="c"></subfield>
        </datafield>

<!--
<parent_blog_name>
Gowers's Weblog
</parent_blog_name>
<parent_blog_uri>
http://blogforever.eu/blog/gowers/
</parent_blog_uri>
-->
        <datafield tag="336" ind1=" " ind2=" ">
            <subfield code="a">news<!-- a vocabulary and
extraction method must be defined in the associated METS
profile. --></subfield>
        </datafield>
        <datafield tag="653" ind1=" " ind2=" ">
            <subfield code="1">News</subfield>
        </datafield>

        <datafield tag="780" ind1=" " ind2=" ">
            <subfield
code="o">http://blogforever.eu/blog/gowers/2012/05/26/a-look-
at-a-few-tripos-questions-ix/</subfield>
        </datafield>
        <datafield tag="785" ind1=" " ind2=" ">
            <subfield
code="o">http://blogforever.eu/blog/gowers/2012/06/08/how-
should-mathematics-be-taught-to-non-mathematicians/</subfield>
        </datafield>
        <datafield tag="532" ind1=" " ind2=" ">
            <subfield code="0">UTF-8</subfield>
        </datafield>
        <datafield tag="041" ind1=" " ind2=" ">
            <subfield code="a">en</subfield>
        </datafield>
        <datafield tag="270" ind1=" " ind2=" ">
            <subfield code="b">GB</subfield>
        </datafield>
</record>
</xmlData>
</mdWrap>
```

```
</dmdSec>

<amdSec>
<!-- The content of the blog post can be dived into sections
corresponding to different mime types and described using
techMD. -->
<techMD ID="text-html1">
<!-- There must be a corresponding techMD section for each of
the 7 images identified in the post. -->
<mdWrap MDTYPE="PREMIS:OBJECT">
<xmlData>
<object xsi:type="file" xmlns="info:lc/xmlns/premis-v2">
<objectIdentifier>
<objectIdentifierType>
URI
<!-- Possible values: URI, DOI, Handle. However, it is
recommended that we use URI identification wherever possible. -
->
</objectIdentifierType>
<objectIdentifierValue>
http://blogforever.eu/blog/gowers/2012/05/31/epsrc-update-
update
</objectIdentifierValue>
</objectIdentifier>
<!-- <objectCategory>
file -->
<!-- Possible values: bitstream, -->
<!-- </objectCategory> -->
<preservationLevel>
<preservationLevelValue>
file
<!-- Possible values: bitstream, file, representation. -->
</preservationLevelValue>
</preservationLevel>
<objectCharacteristics>
<compositionLevel>
0
<!-- This is zero if the text is not compressed. If it is
compressed then there should be another techMD section
describing the compressed object with level 1. In the current
example nothing is assumed to be compressed. -->
</compositionLevel>
<fixity>
<messageDigestAlgorithm>
messageDigestURI/MD5
<!-- This is the URI of the algorithm location used to
calculate the checksum value for the file that this techMD is
recording.For example, MD5 was given as an example. -->
</messageDigestAlgorithm>
<messageDigest>
<!-- The resulting value of the checksum calculated. This can
be got from the tool JHOVE (see tools listed in associated METS
profile - listed in the mets root element of this example) in
some cases. -->
</messageDigest>
</fixity>
```

```
<size>
80376
<!-- This is the size in bytes. Randomly generated value given
above. -->
</size>
<format>
<!-- It might be possible to get the information for this using
the PRONOM registry via JHOVE technical metadata extractor
(http://hul.harvard.edu/jhove/). -->
<formatDesignation>
<formatName>
text/html
<!-- Use if IANA mime type suggested. Vocabulary vc11 (see
associated METS profile). -->
</formatName>
<formatVersion>
1.0
<!--  must be provided wherever possible. -->
</formatVersion>
</formatDesignation>
<formatRegistry>
<formatRegistryName>
http://www.nationalarchives.gov.uk/PRONOM
<!-- For example PRONOM
(http://www.nationalarchives.gov.uk/PRONOM/Default.aspx) and/or
UDFR (http://www.udfr.org/). -->
</formatRegistryName>
<formatRegistryKey>
fmt/102
<!-- This ID is for html 1.0. Not necessarily the correct
version for the blog post being described here. -->
</formatRegistryKey>
</formatRegistry>
</format>
<creatingApplication>
<!-- Information about the blogging software. -->
<creatingApplicationName>
WordPress
</creatingApplicationName>
<creatingApplicationVersion>
unknown
</creatingApplicationVersion>
<dateCreatedByApplication>
2012-05-31T00:00:00</dateCreatedByApplication>
</creatingApplication>
</objectCharacteristics>
<storage>
<contentLocation>
<contentLocationType>
local_directory
</contentLocationType>
<contentLocationValue>
/hermes/archives/blogs/gowers/2012/05/31/epsrc-update-update/
<!-- Offline location on a computer called "hermes" -->
</contentLocationValue>
</contentLocation>
```

```
</storage>
<environment>
<software>
<!-- Requirements for the software rendering the information. I
am not sure how to handle this at the moment. I think this
should list all the browsers on which the repository
implementation is tested on. -->
<swName>
Mozilla Firefox
<!-- This is just an example. -->
</swName>
<swVersion>
</swVersion>
<swType>
</swType>
</software>
<hardware>
<!-- Hardware required to run the software for rendering the
information. I am not qualified to give this information. This
I think would also depend on which equipment we test it on. -->
<hwName>
Ubuntu 12.04
</hwName>
<hwType>
Computer System OS
</hwType>
</hardware>
</environment>

<signatureInformation>
<!-- Only if a signature exists. Should the repository create
one when material is ingested into the repository? -->
<signature>
<signatureEncoding>
</signatureEncoding>
<signatureMethod>
</signatureMethod>
<signatureValue>
</signatureValue>
<signatureValidationRules>
</signatureValidationRules>
</signature>
</signatureInformation>

<linkingEventIdentifier>
    <linkingEventIdentifierType></linkingEventIdentifierType>

<linkingEventIdentifierValue>ingestion1</linkingEventIdentifier
Value>
<!--  If any event exists. -->
</linkingEventIdentifier>
<linkingRightsStatementIdentifier>

<linkingRightsStatementIdentifierType></linkingRightsStatementI
dentifierType>
```

```
<linkingRightsStatementIdentifierValue>rightMD1</linkingRightsS
tatementIdentifierValue>
</linkingRightsStatementIdentifier>
</object>
</xmlData>
</mdWrap>
</techMD>


<techMD ID="post_snapshot_master">
<mdWrap MDTYPE="PREMIS">
<xmlData>
<object xsi:type="file" xmlns="info:lc/xmlns/premis-v2">
<objectIdentifier>
<objectIdentifierType>
URI
<!-- Possible values: URI, DOI, Handle. However, it is
recommended that we use URI identification wherever possible. -
->
</objectIdentifierType>
<objectIdentifierValue>
http://blogforever.eu/blog/gowers/2012/05/31/epsrc-update-
update/masterImage.tif
<!-- The URI of the blog in the repository corresponding to the
screenshot. This is a reference copy of the screenshot. -->
</objectIdentifierValue>
</objectIdentifier>
<!-- <objectCategory>
file -->
<!-- Possible values: bitstream, file, representation. -->
<!-- </objectCategory> -->
<preservationLevel>
<preservationLevelValue>
file
<!-- Possible values: bitstream, file, representation. -->
</preservationLevelValue>
</preservationLevel>
<objectCharacteristics>
<compositionLevel>
0
<!-- This is zero if the screenshot is not compressed. If it is
compressed then there should be another techMD section
describing the compressed object with level 1. In the current
example the screenshot assumed not to be compressed. -->
</compositionLevel>
<fixity>
<messageDigestAlgorithm>
MD5
<!-- This is the URI of the algorithm location used to
calculate the checksum value for the file that this techMD is
recording.For example, MD5 was given as an example. -->
</messageDigestAlgorithm>
<messageDigest>
<!-- The resulting value of the checksum calculated. -->
</messageDigest>
```

```
</fixity>
<size>
97376
<!-- This is the size in bytes. Incorrect value given above. --
>
</size>
<format>
<!-- It might be possible to get the information for this using
the PRONOM registry via JHOVE technical metadata extractor
(http://hul.harvard.edu/jhove/). -->
<formatDesignation>
<formatName>
image/tiff
<!-- Use if IANA mime type suggested. Vocabulary vc11 (see
associated METS profile). -->
</formatName>
<formatVersion>
6.0
<!--  must be provided wherever possible. -->
</formatVersion>
</formatDesignation>
<formatRegistry>
<formatRegistryName>
http://www.nationalarchives.gov.uk/PRONOM
<!-- For example PRONOM
(http://www.nationalarchives.gov.uk/PRONOM/Default.aspx) and/or
UDFR (http://www.udfr.org/). -->
</formatRegistryName>
<formatRegistryKey>
fmt/353
</formatRegistryKey>
</formatRegistry>
</format>
<creatingApplication>
<!-- Information about the tool that created this screenshot. -
->
<creatingApplicationName>
gimp
<!--  This is just an example -->
</creatingApplicationName>
<creatingApplicationVersion>
unknown
</creatingApplicationVersion>
<dateCreatedByApplication>
2007-09-06T00:00:00
</dateCreatedByApplication>
</creatingApplication>
</objectCharacteristics>
<storage>
<contentLocation>
<contentLocationType>
URI
</contentLocationType>
<contentLocationValue>
/hermes/archives/blogs/gowers/2012/05/31/epsrc-update-
update/masterImage.tif
```

```
<!-- Online at the above URI -->
</contentLocationValue>
</contentLocation>
</storage>
<environment>
<software>
<!-- Requirements for the software rendering the information.
not sure how to express this. -->
<swName>
gimp
<!-- This is just an example. -->
</swName>
<swVersion>
2.6.12
</swVersion>
<swType>
</swType>
</software>
<hardware>
<!-- Hardware required to run the software for rendering the
information. -->
<hwName>
Ubuntu 12.04
</hwName>
<hwType>
Computer System OS
</hwType>
</hardware>
</environment>

<signatureInformation>
<!-- Only if a signature exists. Should the repository create
one when material is ingested into the repository? -->
<signature>
<signatureEncoding>
</signatureEncoding>
<signatureMethod>
</signatureMethod>
<signatureValue>
</signatureValue>
<signatureValidationRules>
</signatureValidationRules>
</signature>
</signatureInformation>

<linkingEventIdentifier>
    <linkingEventIdentifierType></linkingEventIdentifierType>

<linkingEventIdentifierValue>master_creation</linkingEventIdent
ifierValue>
</linkingEventIdentifier>
<linkingRightsStatementIdentifier>

<linkingRightsStatementIdentifierType></linkingRightsStatementI
dentifierType>
```

```
<linkingRightsStatementIdentifierValue>rightMD1</linkingRightsS
tatementIdentifierValue>
</linkingRightsStatementIdentifier>
</object>
</xmlData>
</mdWrap>
</techMD>

<techMD ID="post_snapshot_reference">
<mdWrap MDTYPE="PREMIS:OBJECT">
<xmlData>
<object xsi:type="file" xmlns="info:lc/xmlns/premis-v2">
<objectIdentifier>
<objectIdentifierType>
URI
<!-- Possible values: URI, DOI, Handle. However, it is
recommended that we use URI identification wherever possible. -
->
</objectIdentifierType>
<objectIdentifierValue>
http://blogforever.eu/blog/gowers/2012/05/31/epsrc-update-
update/referenceImage.jpg
<!-- The URI of the blog in the repository corresponding to the
screenshot. This is a reference copy of the screenshot. -->
</objectIdentifierValue>
</objectIdentifier>
<!-- <objectCategory>
file -->
<!-- Possible values: bitstream, file, representation. -->
<!-- </objectCategory> -->
<preservationLevel>
<preservationLevelValue>
file
<!-- Possible values: bitstream, file, representation. -->
</preservationLevelValue>
</preservationLevel>
<objectCharacteristics>
<compositionLevel>
0
<!-- This is zero if the screenshot is not compressed. If it is
compressed then there should be another techMD section
describing the compressed object with level 1. In the current
example the screenshot assumed not to be compressed. -->
</compositionLevel>
<fixity>
<messageDigestAlgorithm>
MD5
<!-- This is the URI of the algorithm location used to
calculate the checksum value for the file that this techMD is
recording.For example, MD5 was given as an example. -->
</messageDigestAlgorithm>
<messageDigest>
<!-- The resulting value of the checksum calculated. -->
</messageDigest>
</fixity>
```

```
<size>
97376
<!-- This is the size in bytes. Incorrect value given above. --
>
</size>
<format>
<!-- It might be possible to get the information for this using
the PRONOM registry via JHOVE technical metadata extractor
(http://hul.harvard.edu/jhove/). -->
<formatDesignation>
<formatName>
image/jpeg
<!-- Use if IANA mime type suggested. Vocabulary vc11 (see
associated METS profile). -->
</formatName>
<formatVersion>
2.2
<!--  must be provided wherever possible. -->
</formatVersion>
</formatDesignation>
<formatRegistry>
<formatRegistryName>
http://www.nationalarchives.gov.uk/PRONOM
<!-- For example PRONOM
(http://www.nationalarchives.gov.uk/PRONOM/Default.aspx) and/or
UDFR (http://www.udfr.org/). -->
</formatRegistryName>
<formatRegistryKey>
x-fmt/391
</formatRegistryKey>
</formatRegistry>
</format>
<creatingApplication>
<!-- Information about the tool that created this screenshot. -
->
<creatingApplicationName>
gimp
<!--  This is just an example -->
</creatingApplicationName>
<creatingApplicationVersion>
unknown
</creatingApplicationVersion>
<dateCreatedByApplication>
2007-09-06T00:00:00
</dateCreatedByApplication>
</creatingApplication>
</objectCharacteristics>
<storage>
<contentLocation>
<contentLocationType>
URI
</contentLocationType>
<contentLocationValue>
/hermes/archives/blogs/gowers/2012/05/31/epsrc-update-
update/masterImage.jpg
<!-- Local computer database -->
```

```
</contentLocationValue>
</contentLocation>
</storage>
<environment>
<software>
<!-- Requirements for the software rendering the information.
not sure how to express this. -->
<swName>
gimp
<!-- This is just an example. -->
</swName>
<swVersion>
2.6.12
</swVersion>
<swType>
</swType>
</software>
<hardware>
<!-- Hardware required to run the software for rendering the
information. -->
<hwName>
Ubuntu 12.04
</hwName>
<hwType>
Computer System OS
</hwType>
</hardware>
</environment>

<signatureInformation>
<!-- Only if a signature exists. Should the repository create
one when material is ingested into the repository? -->
<signature>
<signatureEncoding>
</signatureEncoding>
<signatureMethod>
</signatureMethod>
<signatureValue>
</signatureValue>
<signatureValidationRules>
</signatureValidationRules>
</signature>
</signatureInformation>

<linkingEventIdentifier>
    <linkingEventIdentifierType></linkingEventIdentifierType>

<linkingEventIdentifierValue>reference_creation</linkingEventId
entifierValue>
</linkingEventIdentifier>
<linkingRightsStatementIdentifier>

<linkingRightsStatementIdentifierType></linkingRightsStatementI
dentifierType>
```

```
<linkingRightsStatementIdentifierValue>rightMD1</linkingRightsS
tatementIdentifierValue>
</linkingRightsStatementIdentifier>
</object>
</xmlData>
</mdWrap>
</techMD>

<techMD ID="image1">
<!-- A techMD section like this must be created for every image
identified in the post - for both syndicated images and images
belonging to the blog. The images may not be kept in the
repository if not belonging to the blog. For the current
example, we will limit the image to this one example. In the
case of this post, most images are related to gravatars and
other external images. -->
<mdWrap MDTYPE="PREMIS:OBJECT">
<xmlData>
<object xsi:type="file" xmlns="info:lc/xmlns/premis-v2">
<objectIdentifier>
<objectIdentifierType>
URI
<!-- Possible values: URI, DOI, Handle. However, it is
recommended that we use URI identification wherever possible. -
->
</objectIdentifierType>
<objectIdentifierValue>
"http://blogforever.eu/blog/gowers/2012/05/31/epsrc-update-
update/www.gravatar.com/avatar/ad516503a11cd5ca435acc9bb6523536
?s=25&amp;forcedefault=1&amp;d=identicon"
<!-- The URI of the blog in the repository corresponding to the
screenshot. This is a reference copy of the screenshot. -->
</objectIdentifierValue>
</objectIdentifier>
<!-- <objectCategory>
file -->
<!-- Possible values: bitstream, file, representation. -->
<!-- </objectCategory> -->
<preservationLevel>
<preservationLevelValue>
file
<!-- Possible values: bitstream, file, representation. -->
</preservationLevelValue>
</preservationLevel>
<objectCharacteristics>
<compositionLevel>
0
<!-- This is zero if the screenshot is not compressed. If it is
compressed then there should be another techMD section
describing the compressed object with level 1. In the current
example the screenshot assumed not to be compressed. -->
</compositionLevel>
<fixity>
<messageDigestAlgorithm>
MD5
```

```
<!-- This is the URI of the algorithm location used to
calculate the checksum value for the file that this techMD is
recording.For example, MD5 was given as an example. -->
</messageDigestAlgorithm>
<messageDigest>
<!-- The resulting value of the checksum calculated. -->
</messageDigest>
</fixity>
<size>
97376
<!-- This is the size in bytes. Incorrect value given above. --
>
</size>
<format>
<!-- It might be possible to get the information for this using
the PRONOM registry via JHOVE technical metadata extractor
(http://hul.harvard.edu/jhove/). -->
<formatDesignation>
<formatName>
image/png
<!-- Use if IANA mime type suggested. Vocabulary vc11 (see
associated METS profile). -->
</formatName>
<formatVersion>
1.2
<!--  must be provided wherever possible. The one provided here
is fictitious. -->
</formatVersion>
</formatDesignation>
<formatRegistry>
<formatRegistryName>
http://www.nationalarchives.gov.uk/PRONOM
<!-- For example PRONOM
(http://www.nationalarchives.gov.uk/PRONOM/Default.aspx) and/or
UDFR (http://www.udfr.org/). -->
</formatRegistryName>
<formatRegistryKey>
fmt/13
<!--  This id is for Portable Network Graphics version 1.2. -->
</formatRegistryKey>
</formatRegistry>
</format>
<creatingApplication>
<!-- Information about the tool that created this screenshot. -
->
<creatingApplicationName>
gimp
<!--  This is just an example. Could not find the information
with the image. -->
</creatingApplicationName>
<creatingApplicationVersion>
unknown
</creatingApplicationVersion>
<dateCreatedByApplication>
2007-09-06T00:00:00
</dateCreatedByApplication>
```

```
</creatingApplication>
</objectCharacteristics>
<storage>
<contentLocation>
<contentLocationType>
URI
</contentLocationType>
<contentLocationValue>
/hermes/archives/blogs/gowers/2012/05/31/epsrc-update-
update/www.gravatar.com/avatar/ad516503a11cd5ca435acc9bb6523536
?s=25&amp;forcedefault=1&amp;d=identicon
<!-- In this local computer. -->
</contentLocationValue>
</contentLocation>
</storage>
<environment>
<software>
<!-- Requirements for the software rendering the information.
not sure how to express this. -->
<swName>
gimp
<!-- This is just an example. -->
</swName>
<swVersion>
2.6.12
</swVersion>
<swType>
</swType>
</software>
<hardware>
<!-- Hardware required to run the software for rendering the
information. -->
<hwName>
Ubuntu 12.04
</hwName>
<hwType>
Computer System OS
</hwType>
</hardware>
</environment>

<signatureInformation>
<!-- Only if a signature exists. Should the repository create
one when material is ingested into the repository? -->
<signature>
<signatureEncoding>
</signatureEncoding>
<signatureMethod>
</signatureMethod>
<signatureValue>
</signatureValue>
<signatureValidationRules>
</signatureValidationRules>
</signature>
</signatureInformation>
```

```
<linkingEventIdentifier>
    <linkingEventIdentifierType></linkingEventIdentifierType>
    <linkingEventIdentifierValue></linkingEventIdentifierValue>
<!-- If any event should involve this image. -->
</linkingEventIdentifier>
<linkingRightsStatementIdentifier>

<linkingRightsStatementIdentifierType></linkingRightsStatementI
dentifierType>

<linkingRightsStatementIdentifierValue></linkingRightsStatement
IdentifierValue>
<!-- Not sure to whom the rights of a gravatar should belong.
-->
</linkingRightsStatementIdentifier>
</object>
</xmlData>
</mdWrap>
</techMD>

<techMD ID="linkedPage1">
<!-- A techMD section like this must be created for every
hyperlink identified in the post - for both links outside the
blog and within the blog.The target of the links may not be
kept in the repository if not belonging to the blog. For the
current example, we will limit the link to two examples - one
external and one internal. -->
<mdWrap MDTYPE="PREMIS:OBJECT">
<xmlData>
<object xsi:type="file" xmlns="info:lc/xmlns/premis-v2">
<objectIdentifier>
<objectIdentifierType>
URI
<!-- Possible values: URI, DOI, Handle. However, it is
recommended that we use URI identification wherever possible. -
->
</objectIdentifierType>
<objectIdentifierValue>
http://blogforever.eu/blog/gowers/2012/04/13/a-brief-epsrc-
update/
<!-- URI of the target in the repository if the target exists
in the repository. Otherwise the source URI. -->
</objectIdentifierValue>
</objectIdentifier>
<!-- <objectCategory>
file -->
<!-- Possible values: bitstream, file, representation. -->
<!-- </objectCategory> -->
<preservationLevel>
<preservationLevelValue>
file
<!-- Possible values: bitstream, file, representation. -->
</preservationLevelValue>
</preservationLevel>
<objectCharacteristics>
<compositionLevel>
```

```
0
<!-- This is zero if the screenshot is not compressed. If it is
compressed then there should be another techMD section
describing the compressed object with level 1. In the current
example the screenshot assumed not to be compressed. -->
</compositionLevel>
<fixity>
<messageDigestAlgorithm>
MD5
<!-- This is the URI of the algorithm location used to
calculate the checksum value for the file that this techMD is
recording.For example, MD5 was given as an example. -->
</messageDigestAlgorithm>
<messageDigest>
<!-- The resulting value of the checksum calculated. -->
</messageDigest>
</fixity>
<size>
97376
<!-- This is the size in bytes. Incorrect value given above. --
>
</size>
<format>
<!-- It might be possible to get the information for this using
the PRONOM registry via JHOVE technical metadata extractor
(http://hul.harvard.edu/jhove/). -->
<formatDesignation>
<formatName>
text/html
<!-- Use if IANA mime type suggested. Vocabulary vc11 (see
associated METS profile). -->
</formatName>
<formatVersion>
1.0
<!--  must be provided wherever possible. The one provided here
is fictitious. -->
</formatVersion>
</formatDesignation>
<formatRegistry>
<formatRegistryName>
PRONOM
</formatRegistryName>
<formatRegistryKey>
fmt/102
</formatRegistryKey>
</formatRegistry>
</format>
<creatingApplication>
<creatingApplicationName>
</creatingApplicationName>
<creatingApplicationVersion>
</creatingApplicationVersion>
<dateCreatedByApplication>
2012-04-13T10:27:00
</dateCreatedByApplication>
</creatingApplication>
```

```
</objectCharacteristics>
<storage>
<contentLocation>
<contentLocationType>
URI
</contentLocationType>
<contentLocationValue>
/hermes/archives/blogs/gowers/2012/04/13/a-brief-epsrc-update/
<!-- In a local computer. -->
</contentLocationValue>
</contentLocation>
</storage>
<environment>
<software>
<!-- Requirements for the software rendering the information.
not sure how to express this. -->
<swName>
Mozilla Firefox
<!-- This is just an example. -->
</swName>
<swVersion>
13.0
</swVersion>
<swType>
</swType>
</software>
<hardware>
<!-- Hardware required to run the software for rendering the
information. -->
<hwName>
Ubuntu 12.04
</hwName>
<hwType>
Computer System OS
</hwType>
</hardware>
</environment>

<signatureInformation>
<!-- Only if a signature exists. Should the repository create
one when material is ingested into the repository? -->
<signature>
<signatureEncoding>
</signatureEncoding>
<signatureMethod>
</signatureMethod>
<signatureValue>
</signatureValue>
<signatureValidationRules>
</signatureValidationRules>
</signature>
</signatureInformation>

<linkingEventIdentifier>
    <linkingEventIdentifierType></linkingEventIdentifierType>
```

```
<linkingEventIdentifierValue>ingestion2</linkingEventIdentifier
Value>
</linkingEventIdentifier>
<linkingRightsStatementIdentifier>

<linkingRightsStatementIdentifierType></linkingRightsStatementI
dentifierType>

<linkingRightsStatementIdentifierValue>rightsMD1</linkingRights
StatementIdentifierValue>
<!-- Not sure to whom the rights of a gravatar should belong.
-->
</linkingRightsStatementIdentifier>
</object>

</xmlData>
</mdWrap>
</techMD>

<techMD ID="linkedPage2">
<!-- A techMD section like this must be created for every
hyperlink identified in the post - for both links outside the
blog and within the blog.The target of the links may not be
kept in the repository if not belonging to the blog. For the
current example, we will limit the link to two examples - one
external and one internal. -->
<mdWrap MDTYPE="PREMIS:OBJECT">
<xmlData>
<object xsi:type="file" xmlns="info:lc/xmlns/premis-v2">
<objectIdentifier>
<objectIdentifierType>
URI
<!-- Possible values: URI, DOI, Handle. However, it is
recommended that we use URI identification wherever possible. -
->
</objectIdentifierType>
<objectIdentifierValue>
http://www.epsrc.ac.uk/funding/fellows/Pages/areas.aspx
<!-- URI of the target in the repository if the target exists
in the repository. Otherwise the source URI. -->
</objectIdentifierValue>
</objectIdentifier>
<!-- <objectCategory>
file -->
<!-- Possible values: bitstream, file, representation. -->
<!-- </objectCategory> -->
<preservationLevel>
<preservationLevelValue>
reference
<!-- Possible values: reference, bitstream, file,
representation. -->
</preservationLevelValue>
</preservationLevel>
<objectCharacteristics>
<compositionLevel>
```

```
0
<!-- This is zero if the screenshot is not compressed. If it is
compressed then there should be another techMD section
describing the compressed object with level 1. In the current
example the screenshot assumed not to be compressed. -->
</compositionLevel>
<size>
97376
<!-- This is the size in bytes. Incorrect value given above. --
>
</size>
<format>
<!-- It might be possible to get the information for this using
the PRONOM registry via JHOVE technical metadata extractor
(http://hul.harvard.edu/jhove/). -->
<formatDesignation>
<formatName>
text/html
<!-- Use if IANA mime type suggested. Vocabulary vc11 (see
associated METS profile). -->
</formatName>
<formatVersion>
1.0
<!--  must be provided wherever possible. The one provided here
is fictitious. -->
</formatVersion>
</formatDesignation>
<formatRegistry>
<formatRegistryName>
PRONOM
</formatRegistryName>
<formatRegistryKey>
fmt/102
</formatRegistryKey>
</formatRegistry>
</format>
</objectCharacteristics>
<environment>
<software>
<!-- Requirements for the software rendering the information.
not sure how to express this. -->
<swName>
Mozilla Firefox
<!-- This is just an example. -->
</swName>
<swVersion>
13.0
</swVersion>
<swType>
</swType>
</software>
<hardware>
<!-- Hardware required to run the software for rendering the
information. -->
<hwName>
Ubuntu 12.04
```

```
<!-- I think this should be the configuration used to access it
last. -->
</hwName>
<hwType>
Computer System OS
</hwType>
</hardware>
</environment>
</object>
</xmlData>
</mdWrap>
</techMD>

<techMD ID="script1">
<!-- A techMD section like this must be created for every
script identified in the post. For the current example, we will
limit the link to one example. -->
<mdWrap MDTYPE="PREMIS:OBJECT">
<xmlData>
<object xsi:type="file" xmlns="info:lc/xmlns/premis-v2">
<objectIdentifier>
<objectIdentifierType>
URI
<!-- Possible values: URI, DOI, Handle. However, it is
recommended that we use URI identification wherever possible. -
->
</objectIdentifierType>
<objectIdentifierValue>
http://s.stats.wordpress.com/w.js
<!-- URI of the target in the repository if the target exists
in the repository. Otherwise the source URI. -->
</objectIdentifierValue>
</objectIdentifier>
<!-- <objectCategory>
file -->
<!-- Possible values: bitstream, file, representation. -->
<!-- </objectCategory> -->
<preservationLevel>
<preservationLevelValue>
file
<!-- Possible values: reference, bitstream, file,
representation. If the file is unretrievable then the reference
must be kept. -->
</preservationLevelValue>
</preservationLevel>
<objectCharacteristics>
<compositionLevel>
0
<!-- This is zero if the file is not compressed. If it is
compressed then there should be another techMD section
describing the compressed object with level 1. In the current
example all files are assumed not to be compressed. -->
</compositionLevel>
<fixity>
<messageDigestAlgorithm>
MD5
```

```
<!-- This is the URI of the algorithm location used to
calculate the checksum value for the file that this techMD is
recording.For example, MD5 was given as an example. -->
</messageDigestAlgorithm>
<messageDigest>
<!-- The resulting value of the checksum calculated. -->
</messageDigest>
</fixity>
<size>
97376
<!-- This is the size in bytes. Incorrect value given above. --
>
</size>
<format>
<!-- It might be possible to get the information for this using
the PRONOM registry via JHOVE technical metadata extractor
(http://hul.harvard.edu/jhove/). -->
<formatDesignation>
<formatName>
application/javascript
<!-- Use if IANA mime type suggested. Vocabulary vc11 (see
associated METS profile). -->
</formatName>
<formatVersion>
<!--  must be provided wherever possible. The one provided here
is fictitious. -->
</formatVersion>
</formatDesignation>
<formatRegistry>
<formatRegistryName>
PRONOM
</formatRegistryName>
<formatRegistryKey>
x-fmt/423
</formatRegistryKey>
</formatRegistry>
</format>
</objectCharacteristics>
</object>
</xmlData>
</mdWrap>
</techMD>

<techMD ID="feed1">

<!-- A techMD section like this must be created for every feed
identified in the post. For the current example, we will limit
the feed to the blog rss feed. -->
<mdWrap MDTYPE="PREMIS:RIGHTS">
<xmlData>
<object xsi:type="file" xmlns="info:lc/xmlns/premis-v2">
<objectIdentifier>
<objectIdentifierType>
URI
```

```
<!-- Possible values: URI, DOI, Handle. However, it is
recommended that we use URI identification wherever possible. -
->
</objectIdentifierType>
<objectIdentifierValue>
http://blogforever.eu/blog/gowers/feed/
<!-- URI of the target in the repository if the target exists
in the repository. Otherwise the source URI. -->
</objectIdentifierValue>
</objectIdentifier>
<!-- <objectCategory>
file -->
<!-- Possible values: bitstream, file, representation. -->
<!-- </objectCategory> -->
<preservationLevel>
<preservationLevelValue>
file
<!-- Possible values: reference, bitstream, file,
representation. If the file is unretrievable then the reference
must be kept. -->
</preservationLevelValue>
</preservationLevel>
<objectCharacteristics>
<compositionLevel>
0
<!-- In the current example all files are assumed not to be
compressed. -->
</compositionLevel>
<fixity>
<messageDigestAlgorithm>
MD5
<!-- This is the URI of the algorithm location used to
calculate the checksum value for the file that this techMD is
recording.For example, MD5 was given as an example. -->
</messageDigestAlgorithm>
<messageDigest>
<!-- The resulting value of the checksum calculated. -->
</messageDigest>
</fixity>
<size>
97376
<!-- This is the size in bytes. Incorrect value given above. --
>
</size>
<format>
<!-- It might be possible to get the information for this using
the PRONOM registry via JHOVE technical metadata extractor
(http://hul.harvard.edu/jhove/). -->
<formatDesignation>
<formatName>
application/rss+xml
<!-- Use if IANA mime type suggested. Vocabulary vc11 (see
associated METS profile). -->
</formatName>
<formatVersion>
```

```
<!--  must be provided wherever possible. The one provided here
is fictitious. -->
</formatVersion>
</formatDesignation>
<formatRegistry>
<formatRegistryName>
PRONOM
</formatRegistryName>
<formatRegistryKey>
fmt/101
</formatRegistryKey>
</formatRegistry>
</format>
</objectCharacteristics>
</object>
</xmlData>
</mdWrap>
</techMD>

<techMD ID="techMD-0001">
    <mdWrap MDTYPE="NISOIMG">
        <xmlData>
            <mix:mix xmlns:mix="http://www.loc.gov/mix/v20"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/mix/v20
http://www.loc.gov/standards/mix/mix20/mix20.xsd">
                <mix:BasicDigitalObjectInformation>
                    <mix:byteOrder>little
endian</mix:byteOrder>
                    <mix:Compression>

<mix:compressionScheme>LZW</mix:compressionScheme>
                    </mix:Compression>
                </mix:BasicDigitalObjectInformation>
                <mix:BasicImageInformation>
                    <mix:BasicImageCharacteristics>
                        <mix:imageWidth>310</mix:imageWidth>
                        <mix:imageHeight>508</mix:imageHeight>
                        <mix:PhotometricInterpretation>
                            <mix:colorSpace>RGB
Palette</mix:colorSpace>
                        </mix:PhotometricInterpretation>
                    </mix:BasicImageCharacteristics>
                </mix:BasicImageInformation>
                <mix:ImageCaptureMetadata>
                    <mix:GeneralCaptureInformation/>
                    <mix:orientation>normal*</mix:orientation>
                </mix:ImageCaptureMetadata>
                <mix:ImageAssessmentMetadata>
                    <mix:SpatialMetrics/>
                    <mix:ImageColorEncoding>
                        <mix:BitsPerSample>

<mix:bitsPerSampleValue>8</mix:bitsPerSampleValue>

<mix:bitsPerSampleUnit>integer</mix:bitsPerSampleUnit>
```

```
                        </mix:BitsPerSample>
                    </mix:ImageColorEncoding>
                </mix:ImageAssessmentMetadata>
            </mix:mix>
        </xmlData>
    </mdWrap>
</techMD>

<techMD ID="techMD-0002">
    <mdWrap MDTYPE="NISOIMG">
        <xmlData>
            <mix:mix xmlns:mix="http://www.loc.gov/mix/v20"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/mix/v20
http://www.loc.gov/standards/mix/mix20/mix20.xsd">
                <mix:BasicDigitalObjectInformation>
                    <mix:byteOrder>big endian</mix:byteOrder>
                    <mix:Compression>
                        <mix:compressionScheme>JPEG (old-
style)</mix:compressionScheme>
                    </mix:Compression>
                </mix:BasicDigitalObjectInformation>
                <mix:BasicImageInformation>
                    <mix:BasicImageCharacteristics>
                        <mix:imageWidth>2048</mix:imageWidth>
                        <mix:imageHeight>1536</mix:imageHeight>
                        <mix:PhotometricInterpretation>
                            <mix:YCbCr>
                                <mix:YCbCrSubSampling>

<mix:yCbCrSubsampleHoriz>2</mix:yCbCrSubsampleHoriz>

<mix:yCbCrSubsampleVert>1</mix:yCbCrSubsampleVert>
                                </mix:YCbCrSubSampling>
                            </mix:YCbCr>
                        </mix:PhotometricInterpretation>
                    </mix:BasicImageCharacteristics>
                </mix:BasicImageInformation>
                <mix:ImageCaptureMetadata>
                    <mix:GeneralCaptureInformation>
                        <mix:dateTimeCreated>2009-03-
18T13:06:18.0Z</mix:dateTimeCreated>
                        <mix:captureDevice>digital still
camera</mix:captureDevice>
                    </mix:GeneralCaptureInformation>
                    <mix:DigitalCameraCapture>

<mix:digitalCameraManufacturer>Canon</mix:digitalCameraManufact
urer>
                        <mix:DigitalCameraModel>
                            <mix:digitalCameraModelName>Canon
PowerShot SD400</mix:digitalCameraModelName>
                        </mix:DigitalCameraModel>
                        <mix:CameraCaptureSettings>
                            <mix:ImageData>
                                <mix:fNumber>2.8</mix:fNumber>
```

```
<mix:isoSpeedRatings>141</mix:isoSpeedRatings>

<mix:exifVersion>0220</mix:exifVersion>
                                <mix:shutterSpeedValue>

<mix:numerator>1</mix:numerator>

<mix:denominator>8</mix:denominator>
                                </mix:shutterSpeedValue>
                                <mix:apertureValue>

<mix:numerator>280</mix:numerator>

<mix:denominator>100</mix:denominator>
                                </mix:apertureValue>
                                <mix:exposureBiasValue>

<mix:numerator>0</mix:numerator>

<mix:denominator>1</mix:denominator>
                                </mix:exposureBiasValue>
                                <mix:maxApertureValue>

<mix:numerator>280</mix:numerator>

<mix:denominator>100</mix:denominator>
                                </mix:maxApertureValue>

<mix:meteringMode>Pattern</mix:meteringMode>

<mix:lightSource>unknown</mix:lightSource>
                                <mix:flash>Flash fired, auto
mode, red-eye reduction mode</mix:flash>

<mix:focalLength>5.8</mix:focalLength>
                                <mix:sensingMethod>One-chip
colour area sensor</mix:sensingMethod>
                            </mix:ImageData>
                        </mix:CameraCaptureSettings>
                    </mix:DigitalCameraCapture>
                    <mix:orientation>normal*</mix:orientation>
                </mix:ImageCaptureMetadata>
                <mix:ImageAssessmentMetadata>
                    <mix:SpatialMetrics>

<mix:samplingFrequencyUnit>in.</mix:samplingFrequencyUnit>
                        <mix:xSamplingFrequency>
                            <mix:numerator>180</mix:numerator>

<mix:denominator>1</mix:denominator>
                        </mix:xSamplingFrequency>
                        <mix:ySamplingFrequency>
                            <mix:numerator>180</mix:numerator>

<mix:denominator>1</mix:denominator>
```

```
                        </mix:ySamplingFrequency>
                    </mix:SpatialMetrics>
                    <mix:ImageColorEncoding>

<mix:samplesPerPixel>3</mix:samplesPerPixel>
                    </mix:ImageColorEncoding>
                </mix:ImageAssessmentMetadata>
            </mix:mix>
        </xmlData>
    </mdWrap>
</techMD>

<techMD ID="techMD-0003">
    <mdWrap MDTYPE="OTHER">
        <xmlData>
            <docmd:document
                xmlns:docmd="http://www.fcla.edu/docmd"
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance"
                xsi:schemaLocation="http://www.fcla.edu/docmd
http://www.fcla.edu/dls/md/docmd.xsd">
                <docmd:PageCount>20</docmd:PageCount>
                <docmd:WordCount>4224</docmd:WordCount>

<docmd:CharacterCount>24083</docmd:CharacterCount>
                <docmd:Language>U.S. English</docmd:Language>
            </docmd:document>
        </xmlData>
    </mdWrap>
</techMD>

<techMD ID="techMD-0004">
    <mdWrap MDTYPE="TEXTMD">
        <xmlData>
            <textMD:textMD xmlns:textMD="info:lc/xmlns/textMD-
v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="info:lc/xmlns/textMD-v3
http://www.loc.gov/standards/textMD/textMD-v3.01a.xsd">
                <textMD:character_info>
                    <textMD:charset>UTF-8</textMD:charset>
                </textMD:character_info>
                <textMD:markup_basis
version="1.0">HTML</textMD:markup_basis>
            </textMD:textMD>
        </xmlData>
    </mdWrap>
</techMD>

<techMD ID="techMD-0005">
    <mdWrap MDTYPE="OTHER">
        <xmlData>
            <docmd:document
xmlns:docmd="http://www.fcla.edu/docmd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.fcla.edu/docmd
http://www.fcla.edu/dls/md/docmd.xsd">
```

```
                <docmd:PageCount>13</docmd:PageCount>
            </docmd:document>
        </xmlData>
    </mdWrap>
</techMD>

<techMD ID="techMD-0006">
    <mdWrap MDTYPE="OTHER">
        <xmlData>
            <aes:audioObject
xmlns:aes="http://www.aes.org/audioObject"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.aes.org/audioObject
http://www.aes.org/standards/schemas/aes57-2011-08-27.xsd"
ID="AUDIO_OBJECT_00e990cc-b222-42ed-981d-1829cc13062c"
analogDigitalFlag="FILE_DIGITAL" schemaVersion="1.0.0"
disposition="">
                <aes:format specificationVersion="1">MPEG 1/2
Audio Layer 3</aes:format>
                <aes:use useType="OTHER" otherType="unknown"/>
                <aes:primaryIdentifier
identifierType="FILE_NAME">podcast.mp3</aes:primaryIdentifier>
                <aes:face
audioObjectRef="AUDIO_OBJECT_00e990cc-b222-42ed-981d-
1829cc13062c" direction="NONE" ID="FACE_2d32c81e-556e-4d59-
b9a5-79972461d21e" label="face 1">
                    <aes:timeline>
                        <aes:startTime
editRate="1">0</aes:startTime>
                        <aes:duration
editRate="1000">1145212</aes:duration>
                    </aes:timeline>
                    <aes:region
formatRef="FORMAT_REGION_bad68c85-b81e-43f8-a92e-f9b17a50dc58"
ID="REGION_344e104f-38de-4d8a-8d58-c40dc8b35e77" label="region
1" faceRef="FACE_2d32c81e-556e-4d59-b9a5-79972461d21e">
                        <aes:timeRange>
                            <aes:startTime
editRate="1">0</aes:startTime>
                            <aes:duration
editRate="1000">1145212</aes:duration>
                        </aes:timeRange>
                        <aes:numChannels>1</aes:numChannels>
                        <aes:stream ID="STREAM_887692ad-dd8d-
45d7-9b9b-90c50f2cd55c" label="stream 0"
faceRegionRef="REGION_344e104f-38de-4d8a-8d58-c40dc8b35e77">
                            <aes:channelAssignment
frontRearPosition="0.0" channelNum="0"
leftRightPosition="0.0"/>
                        </aes:stream>
                    </aes:region>
                </aes:face>
                <aes:formatList>
                    <aes:formatRegion
ID="FORMAT_REGION_bad68c85-b81e-43f8-a92e-f9b17a50dc58"
```

```
ownerRef="REGION_344e104f-38de-4d8a-8d58-c40dc8b35e77"
label="format region 1" xsi:type="aes:formatRegionType">

<aes:sampleRate>44100.0</aes:sampleRate>
                        <aes:soundField>MONO</aes:soundField>
                        <aes:bitrateReduction>
                            <aes:codecName/>
                            <aes:codecNameVersion/>
                            <aes:codecCreatorApplication/>

<aes:codecCreatorApplicationVersion/>

<aes:codecQuality>LOSSY</aes:codecQuality>
                            <aes:dataRate>64000</aes:dataRate>

<aes:dataRateMode>FIXED</aes:dataRateMode>
                        </aes:bitrateReduction>
                    </aes:formatRegion>
                </aes:formatList>
            </aes:audioObject>
        </xmlData>
    </mdWrap>
</techMD>

<rightsMD ID="rightsMD1">
   <mdWrap MDTYPE="PREMIS:RIGHTS">
       <xmlData>
           <rights xmlns="info:lc/xmlns/premis-v2">

                        <!--
Rights metadata goes here for access, modification, copy,
distribution, licenses.
 -->
                        <rightsStatement>
                            <rightsStatementIdentifier>
                                <rightsStatementIdentifierType>
                                    URI

</rightsStatementIdentifierType>

<rightsStatementIdentifierValue>

http://blogforever.eu/blog/gowers/rights.html

</rightsStatementIdentifierValue>
                            </rightsStatementIdentifier>
                            <rightsBasis>
                                copyright
                                <!--  Is it about copyright,
license, or statute? -->
                            </rightsBasis>
                            <copyrightInformation>
                                <copyrightStatus>
                                    copyrighted
                                    <!-- whether it is
copyrighted, publicdomain or unknown. -->
```

```
                                </copyrightStatus>
                                <copyrightJurisdiction>
                                    GB
                                    <!-- country of
applicability. -->
                                </copyrightJurisdiction>

<copyrightStatusDeterminationDate>
                                    2007-09-07T00:00:00
                                    <!-- when it was
determined, or will be determined -->

</copyrightStatusDeterminationDate>
                                <copyrightNote>
                                    Copyrighted to creator at
creation.
                                    <!-- validity period.
forexample expiration date -->
                                </copyrightNote>
                            </copyrightInformation>
                            <rightsGranted>
                                <!-- If any rights are granted
to replicate, modify, migrate, use, desseminate, or delete the
resource, then it should be indicated here. -->
                                <act>
                                    fair use
                                </act>
                                <restriction>
                                    <!-- For example, "no more
than three copies". -->
                                </restriction>
                                <termOfGrant>
                                    <startDate>
                                        2007-09-07T00:00:00
                                    </startDate>
                                    <!-- <endDate>
                                        Until further notice
                                    </endDate> -->
                                </termOfGrant>
                                <rightsGrantedNote>
                                    Must be taken down if the
creator of the blog requests its deletion.
                                    <!-- For example, when it
is not clear what rights are granted -->
                                </rightsGrantedNote>
                            </rightsGranted>
                            <linkingAgentIdentifier>
                                <linkingAgentIdentifierType>
                                    URI
                                </linkingAgentIdentifierType>
                                <linkingAgentIdentifierValue>
                                    http://gravatar.com/gowers
                                </linkingAgentIdentifierValue>
                            </linkingAgentIdentifier>
                            <linkingAgentIdentifier>
                                <linkingAgentIdentifierType>
```

```
                                    URI
                                </linkingAgentIdentifierType>
                                <linkingAgentIdentifierValue>

http://blogforever.eu/blog/gowers
                                    <!-- URI of the blog in the
repository -->
                                </linkingAgentIdentifierValue>
                            </linkingAgentIdentifier>
                        </rightsStatement>
                    </rights>
        </xmlData>
    </mdWrap>
</rightsMD>

<rightsMD ID="rightMD2">
    <!-- This describes the agent responsible for creating the
blog as described in "rightMD1". -->
    <mdWrap MDTYPE="PREMIS:AGENT">
        <xmlData>
           <agent xmlns="info:lc/xmlns/premis-v2">
                        <agentIdentifier>
                            <agentIdentifierType>
                                URI
                            </agentIdentifierType>
                            <agentIdentifierValue>
                                http://gravatar.com/gowers
                            </agentIdentifierValue>
                        </agentIdentifier>
                        <agentName>
                            Tim Gowers
                        </agentName>
                        <agentType>
                            Person
                        </agentType>
                        <agentNote>
                        </agentNote>
                    </agent>
        </xmlData>
    </mdWrap>
</rightsMD>

<digiprovMD ID="digiProvMD0-0">
<!-- This is the record describing the crawl of the blog post.
-->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD0-0
</eventIdentifierValue>
</eventIdentifier>
```

```
<eventType>
capture
<!-- The value for this field should come from the controlled
vocabulary for repository events (see associated METS profile)
-->
</eventType>
<eventDateTime>0001-01-01T00:00:00
</eventDateTime>
<eventDetail>
Crawling of the blog in preparation of ingestion into the
repository.
</eventDetail>
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service/software. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
softwareURI/repository_crawler
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://gowers.wordpress.com/2012/05/31/epsrc-update-update/
<!-- The source URI of the blog. Always the source in this
field of an event identifier. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD0-1">
<!-- This is the record describing the crawl of linkedPage1. --
>
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD0-1
</eventIdentifierValue>
</eventIdentifier>
<eventType>
capture
```

```
<!-- The value for this field should come from the controlled
vocabulary for repository events (see associated METS profile)
-->
</eventType>
    <eventDateTime>0001-01-01T00:00:00
</eventDateTime>
<eventDetail>
Crawling of the blog in preparation of ingestion into the
repository.
</eventDetail>
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service/software. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
softwareURI/repository_crawler
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://gowers.wordpress.com/2012/04/13/a-brief-epsrc-update/
<!-- The source URI of the blog. Always the source in this
field of an event identifier. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>

</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD0-2">
<!-- This is the record describing the crawl of script1. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD0-2
</eventIdentifierValue>
</eventIdentifier>
<eventType>
capture
<!-- The value for this field should come from the controlled
vocabulary for repository events (see associated METS profile)
-->
```

```
</eventType>
    <eventDateTime>0001-01-01T00:00:00
</eventDateTime>
<eventDetail>
Crawling of the blog in preparation of ingestion into the
repository.
</eventDetail>
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service/software. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
softwareURI/repository_crawler
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://s.stats.wordpress.com/w.js
<!-- The source URI of the blog. Always the source in this
field of an event identifier. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD0-3">
<!-- This is the record describing the crawl of feed1. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD0-3
</eventIdentifierValue>
</eventIdentifier>
<eventType>
capture
<!-- The value for this field should come from the controlled
vocabulary for repository events (see associated METS profile)
-->
</eventType>
    <eventDateTime>0001-01-01T00:00:00
</eventDateTime>
<eventDetail>
```

```
Crawling of the blog in preparation of ingestion into the
repository.
</eventDetail>
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service/software. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
softwareURI/repository_crawler
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://gowers.wordpress.com/feed/
<!-- The source URI of the blog. Always the source in this
field of an event identifier. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD1">
<!-- This section describes the agent responsible for the crawl
described in "digiProvMD0-0", "digiProvMD0-1", "digiProvMD0-2",
"digiProvMD0-3". -->
<mdWrap MDTYPE="PREMIS:AGENT">
<xmlData>
<agent xmlns="info:lc/xmlns/premis-v2">
<agentIdentifier>
<agentIdentifierType>
URI
</agentIdentifierType>
<agentIdentifierValue>
softwareURI/repository_crawler
</agentIdentifierValue>
</agentIdentifier>
<agentName>
BlogForever_spider
</agentName>
<agentType>
service
</agentType>
<agentNote>
</agentNote>
</agent>
</xmlData>
</mdWrap>
```

```
</digiprovMD>

<digiprovMD ID="digiProvMD2-0">
<!-- This section records the event that created the master
screenshot of the blog post. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD2-0
</eventIdentifierValue>
</eventIdentifier>
<eventType>
screenshot_creation
<!-- The value for this field should come from the controlled
vocabulary for repository events (see associated METS profile)
-->
</eventType>
<eventDateTime>
2012-06-06T14:00:00-06:00</eventDateTime>
<eventDetail>
Creation of master screenshot for the blog.
</eventDetail>
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
softwareURI/master_screenshot_software
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://gowers.wordpress.com/2012/05/31/epsrc-update-update/
<!-- The source URI of the blog page from which the screenshot
is being created. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD3">
```

```
<!-- This record describes the agent involved in creating the
master screenshot (the event described in "digiProvMD2-0"). -->
<mdWrap MDTYPE="PREMIS:AGENT">
<xmlData>
<agent xmlns="info:lc/xmlns/premis-v2">
<agentIdentifier>
<agentIdentifierType>
URI
</agentIdentifierType>
<agentIdentifierValue>
softwareURI/master_screenshot_software
</agentIdentifierValue>
</agentIdentifier>
<agentName>
master_screenshot_software
</agentName>
<agentType>
service
</agentType>
<agentNote>
</agentNote>
</agent>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD2-1">
<!-- This section records the event that created the reference
screenshot of the blog post. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD2-1
</eventIdentifierValue>
</eventIdentifier>
<eventType>
screenshot_creation
<!-- The value for this field should come from the controlled
vocabulary for repository events (see associated METS profile)
-->
</eventType>
<eventDateTime>
2012-06-06T14:00:00-06:00</eventDateTime>
<eventDetail>
Creation of reference screenshot for the blog.
</eventDetail>
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service. -->
<linkingAgentIdentifierType>
URI
```

```
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
softwareURI/reference_screenshot_software
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://gowers.wordpress.com/2012/05/31/epsrc-update-update/
<!-- The source URI of the blog page from which the screenshot
is being created. If the reference is created from the master
then the mater URI should go here. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD5">
<!-- This record describes the agent involved in creating the
reference screenshot (the event described in "digiProvMD2-1").
-->
<mdWrap MDTYPE="PREMIS:AGENT">
<xmlData>
<agent xmlns="info:lc/xmlns/premis-v2">
<agentIdentifier>
<agentIdentifierType>
URI
</agentIdentifierType>
<agentIdentifierValue>
softwareURI/reference_screenshot_software
</agentIdentifierValue>
</agentIdentifier>
<agentName>
reference_screenshot_software
</agentName>
<agentType>
service
</agentType>
<agentNote>
</agentNote>
</agent>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD6">
<!-- This section describes the metadata creation event. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
```

```
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD6
</eventIdentifierValue>
</eventIdentifier>
<eventType>
metadata_creation
<!-- The value for this field should come from the controlled
vocabulary for repository events (see associated METS profile)
-->
</eventType>
<eventDateTime>
2012-07-06T12:00:00</eventDateTime>
<eventDetail>
Metadata created for the blog.
</eventDetail>
<!-- <eventOutcomeInformation>
</eventOutcomeInformation> -->
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
mailto:yunhyong.kim@glasgow.ac.uk
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- Repository URI of the blog. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://blogforever.eu/blog/gower/2012/05/31/epsrc-update-
update/
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD7">
<!-- This section describes the agent who created the metadata
in the METS object. -->
<mdWrap MDTYPE="PREMIS:AGENT">
<xmlData>
<agent xmlns="info:lc/xmlns/premis-v2">
<agentIdentifier>
<agentIdentifierType>
URI
</agentIdentifierType>
```

```
<agentIdentifierValue>
mailto:yunhyong.kim@glasgow.ac.uk
</agentIdentifierValue>
</agentIdentifier>
<agentName>
Yunhyong Kim
</agentName>
<agentType>
person
</agentType>
<agentNote>
</agentNote>
</agent>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD8">
<!-- This event decribes the creation of the blog post itself,
before it was ingested into the repository. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD8
</eventIdentifierValue>
</eventIdentifier>
<eventType>
content_creation
<!-- The value for this field should come from the controlled
vocabulary for repository events (see associated METS profile)
-->
</eventType>
    <eventDateTime>0001-01-01T00:00:00
</eventDateTime>
<eventDetail>
Creation of the blog itself.
</eventDetail>
<!-- <eventOutcomeInformation>
</eventOutcomeInformation> -->
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
http://gravatar.com/gowers
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
```

```
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://gowers.wordpress.com/2012/05/31/epsrc-update-update/
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD9">
<!-- This describes the agent responsible for creating the blog
as described in "digiProvMD8". -->
<mdWrap MDTYPE="PREMIS:AGENT">
<xmlData>
<agent xmlns="info:lc/xmlns/premis-v2">
<agentIdentifier>
<agentIdentifierType>
URI
</agentIdentifierType>
<agentIdentifierValue>
http://gravatar.com/gowers
</agentIdentifierValue>
</agentIdentifier>
<agentName>
Tim Gowers
</agentName>
<agentType>
Person
</agentType>
<agentNote>
</agentNote>
</agent>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD10-0">
<!-- This section describes the ingestion of the blog post into
the repository. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD10-0
</eventIdentifierValue>
</eventIdentifier>
<eventType>
```

```
ingestion
<!-- For example, "ingestion" -->
</eventType>
    <eventDateTime>0001-01-01T00:00:00
</eventDateTime>
<eventDetail>
Ingestion of the blog into the repository.
</eventDetail>
<!-- <eventOutcomeInformation>
</eventOutcomeInformation> -->
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
softwareURI/blogforever_repository
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://gowers.wordpress.com/2012/05/31/epsrc-update-update/
<!-- The source URI of the blog post. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD10-1">
<!-- This section describes the ingestion of the linkedPage1
into the repository. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD10-1
</eventIdentifierValue>
</eventIdentifier>
<eventType>
ingestion
<!-- For example, "ingestion" -->
</eventType>
    <eventDateTime>0001-01-01T00:00:00
</eventDateTime>
```

```
<eventDetail>
Ingestion of the blog into the repository.
</eventDetail>
<!-- <eventOutcomeInformation>
</eventOutcomeInformation> -->
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
softwareURI/blogforever_repository
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://gowers.wordpress.com/2012/04/13/a-brief-epsrc-update/
<!-- The source URI of linkedPage1. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD10-2">
<!-- This section describes the ingestion of script1. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD10-2
</eventIdentifierValue>
</eventIdentifier>
<eventType>
ingestion
<!-- For example, "ingestion" -->
</eventType>
    <eventDateTime>0001-01-01T00:00:00
</eventDateTime>
<eventDetail>
Ingestion of the blog into the repository.
</eventDetail>
<!-- <eventOutcomeInformation>
</eventOutcomeInformation> -->
<linkingAgentIdentifier>
```

```
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>
softwareURI/blogforever_repository
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://s.stats.wordpress.com/w.js
<!-- The URI of script1. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD10-3">
<!-- This section describes the ingestion of the feed1 into the
repository. -->
<mdWrap MDTYPE="PREMIS:EVENT">
<xmlData>
<event xmlns="info:lc/xmlns/premis-v2">
<eventIdentifier>
<eventIdentifierType>
Internal_XML_ID
</eventIdentifierType>
<eventIdentifierValue>
ProvMD10-3
</eventIdentifierValue>
</eventIdentifier>
<eventType>
ingestion
<!-- For example, "ingestion" -->
</eventType>
    <eventDateTime>0001-01-01T00:00:00
</eventDateTime>
<eventDetail>
Ingestion of the blog into the repository.
</eventDetail>
<!-- <eventOutcomeInformation>
</eventOutcomeInformation> -->
<linkingAgentIdentifier>
<!-- This links the event to the agent responsible, whether
this is a person, organisation, or service. -->
<linkingAgentIdentifierType>
URI
</linkingAgentIdentifierType>
```

```
<linkingAgentIdentifierValue>
softwareURI/blogforever_repository
</linkingAgentIdentifierValue>
</linkingAgentIdentifier>
<linkingObjectIdentifier>
<!-- In this case this would be the source blog URI to which
this events relates. -->
<linkingObjectIdentifierType>
URI
</linkingObjectIdentifierType>
<linkingObjectIdentifierValue>
http://gowers.wordpress.com/feed/
<!-- The source URI of feed1. -->
</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
</xmlData>
</mdWrap>
</digiprovMD>

<digiprovMD ID="digiProvMD11">
<!-- Record of agents responsible for ingesting the blog post,
linked page, feed, and script into the repository
(digiProvMD10-0, 10-1,10-2,and 10-3). -->
<mdWrap MDTYPE="PREMIS:AGENT">
<xmlData>
<agent xmlns="info:lc/xmlns/premis-v2">
<agentIdentifier>
<agentIdentifierType>
URI
</agentIdentifierType>
<agentIdentifierValue>
softwareURI/blogforever_respository
</agentIdentifierValue>
</agentIdentifier>
<agentName>
blogforever_repository
</agentName>
<agentType>
service
</agentType>
<agentNote>
</agentNote>
</agent>
</xmlData>
</mdWrap>
</digiprovMD>

</amdSec>

<fileSec>
<!-- in this section it is explained how all the files
described with the amistrative and descriptive metadata are
grouped together. In the case of a blog, the only file
associated to the blog in this record is the snapshop.
Although, pdf has been used so far for this purpose, it is
```

```
recommended that a tiff image be created as master snapshot and
jpeg be displayed. -->
<fileGrp>
<fileGrp>
    <file
ID="http__blogforever.eu_blog_gowers_2012_05_31_epsrc-update-
update" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="text/html" CREATED="2012-06-
06T14:26:40-06:00" ADMID="post_snapshot_master">
<!-- This is the master copy blog post html. -->
<FLocat LOCTYPE="URL"
xlink:href="/backupstorageLoc/blog/gowers/2012/05/31/epsrc-
update-update"></FLocat>
</file>
<file ID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-update_masterImage.tif" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="text/tiff" CREATED="2012-06-
06T14:26:40-06:00" ADMID="post_snapshot_master">
<!-- This is a archival quality master screen shot of the blog
post. -->
<FLocat LOCTYPE="URL"
xlink:href="/backupstorageLoc/blog/gowers/2012/05/31/epsrc-
update-update/masterImage.tif"></FLocat>
</file>
<file ID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-
update_www.gravatar.com_avatar_ad516503a11cd5ca435acc9bb6523536
_s_25_forcedefault_1_d_identicon" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="image/png" CREATED="2012-06-
06T14:26:40-06:00" ADMID="post_snapshot_master">
<!-- This is image1. Not stored in the repository. -->
<FLocat LOCTYPE="URL"
xlink:href="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-
update/www.gravatar.com/avatar/ad516503a11cd5ca435acc9bb6523536
?s=25&amp;forcedefault=1&amp;d=identicon"></FLocat>
</file>
<file ID="http___blogforever.eu_blog_gowers_2012_04_13_a-brief-
epsrc-update_" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="text/html" CREATED="2012-06-
06T14:26:40-06:00" ADMID="post_snapshot_master">
<!-- This is linkedPage1 master copy. -->
<FLocat LOCTYPE="URL"
xlink:href="/backupstorageLoc/blog/gowers/2012/04/13/a-brief-
epsrc-update/"></FLocat>
</file>
<file
ID="http___www.epsrc.ac.uk_funding_fellows_Pages_areas.aspx"
SIZE="3126" CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="text/html" CREATED="2012-06-
06T14:26:40-06:00" ADMID="post_snapshot_master">
<!-- This is linkedPage2. Not stored in the repository -->
```

```
<FLocat LOCTYPE="URL"
xlink:href="http://www.epsrc.ac.uk/funding/fellows/Pages/areas.
aspx"></FLocat>
</file>
<file ID="http___s.stats.wordpress.com_w.js" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="application/javascript"
CREATED="2012-06-06T14:26:40-06:00"
ADMID="post_snapshot_master">
<!-- This is script1 master copy. -->
<FLocat LOCTYPE="URL"
xlink:href="/backupstorageLoc/blog/gowers/s.stats.wordpress.com
/w.js"></FLocat>
</file>
<file ID="http___blogforever.eu_blog_gowers_feed_" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="application/rss+xml"
CREATED="2012-06-06T14:26:40-06:00"
ADMID="post_snapshot_master">
<!-- This is feed1 master copy. -->
<FLocat LOCTYPE="URL"
xlink:href="/backupstorageLoc/blog/gowers/feed/"></FLocat>
</file>
</fileGrp>
<fileGrp>
<file ID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-update" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="text/html" CREATED="2012-06-
06T14:26:40-06:00" ADMID="post_snapshot_master">
<!-- This is the reference blog post html. -->
<FLocat LOCTYPE="URL"
xlink:href="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-update"></FLocat>
</file>
<file ID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-update_referenceImage.jpg" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="image/jpeg" CREATED="2012-06-
06T14:26:40-06:00" ADMID="post_snapshot_master">
<!-- This is the reference screen shot of the blog post. -->
<FLocat LOCTYPE="URL"
xlink:href="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-update/masterImage.jpg"></FLocat>
</file>
<file
ID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc_update_u
pdate_www.gravatar.com_avatar_ad516503a11cd5ca435acc9bb6523536_
s_25_forcedefault_1_d_identicon" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="image/png" CREATED="2012-06-
06T14:26:40-06:00" ADMID="post_snapshot_master">
<!-- This is image1. Not stored in the repository. -->
<FLocat LOCTYPE="URL"
xlink:href="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-
```

```
update/www.gravatar.com/avatar/ad516503a11cd5ca435acc9bb6523536
?s=25&amp;forcedefault=1&amp;d=identicon"></FLocat>
</file>
<file
ID="http___blogforever.eu_blog_gowers_2012_04_13_a_brief_epsrc_
update_" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="text/html" CREATED="2012-06-
06T14:26:40-06:00" ADMID="post_snapshot_master">
<!-- This is reference linkedPage1. -->
<FLocat LOCTYPE="URL"
xlink:href="http://blogforever.eu/blog/gowers/2012/04/13/a-
brief-epsrc-update/"></FLocat>
</file>

    <!-- The ID of the following file has already appeared in
the document:
<file
ID="http___www.epsrc.ac.uk_funding_fellows_Pages_areas.aspx"
SIZE="3126" CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="text/html" CREATED="2012-06-
06T14:26:40-06:00" ADMID="techMD2">
// This is linkedPage2. Not stored in the repository. //
<FLocat LOCTYPE="URL"
xlink:href="http://www.epsrc.ac.uk/funding/fellows/Pages/areas.
aspx"></FLocat>
</file>
    -->

    <!-- The ID of the following file has already appeared in
the document:
<file ID="http___s.stats.wordpress.com_w.js" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="application/javascript"
CREATED="2012-06-06T14:26:40-06:00" ADMID="techMD2">
// This is a reference copy of script1. //
<FLocat LOCTYPE="URL"
xlink:href="http://s.stats.wordpress.com/w.js"></FLocat>
</file>
-->

    <!-- The ID of the following file has already appeared in
the document:
<file ID="http___blogforever.eu_blog_gowers_feed_" SIZE="3126"
CHECKSUM="4ea7325ecef266792a03e5f82ce67762970e14a9"
CHECKSUMTYPE="SHA-1" MIMETYPE="application/rss+xml"
CREATED="2012-06-06T14:26:40-06:00" ADMID="techMD2">
// This is a reference copy of feed1. //
<FLocat LOCTYPE="URL"
xlink:href="http://blogforever.eu/blog/gowers/feed/"></FLocat>
</file>
-->
</fileGrp>
<fileGrp USE="fictive example">
```

```
    <!-- The following files are fictive examples to
demonstrate the inclusion of files with different digital
formats and how to associate them with technical metadata. -->
    <file ID="file-0001" MIMETYPE="image/gif" ADMID="techMD-
0001">
        <FLocat LOCTYPE="URL" xlink:href="file:wine_gif.gif"/>
    </file>
    <file ID="file-0002" MIMETYPE="image/jpg" ADMID="techMD-
0002">
        <FLocat LOCTYPE="URL" xlink:href="file:wine.jpg"/>
    </file>
    <file ID="file-0003" MIMETYPE="application/msword"
ADMID="techMD-0003">
        <FLocat LOCTYPE="URL"
xlink:href="file:BlogForever_1st_periodic_report_20120306_UW.do
c"/>
    </file>
    <file ID="file-0004" MIMETYPE="text/html" ADMID="techMD-
0004">
        <FLocat LOCTYPE="URL"  xlink:href="file:Miettes Bedtime
Story Podcast.htm"/>
    </file>
    <file ID="file-0005" MIMETYPE="application/pdf"
ADMID="techMD-0005">
        <FLocat LOCTYPE="URL" xlink:href="file:woolrich.pdf"/>
    </file>
    <file ID="file-0006" MIMETYPE="audio/mpeg3" ADMID="techMD-
0006">
        <FLocat LOCTYPE="URL" xlink:href="file:podcast.mp3"/>
    </file>
</fileGrp>
</fileGrp>
</fileSec>

<structMap>
<div TYPE="WEB_CAPTURE" LABEL="EPSRC update update"
DMDID="dmdMD2">
<fptr
FILEID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-update">
    <area
FILEID="http__blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-update" BEGIN="105" EXTENT="31264" BETYPE="BYTE"
EXTTYPE="BYTE"/>
</fptr>
<fptr
FILEID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-update_referenceImage.jpg">
    <area
FILEID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-update_masterImage.tif" BEGIN="105" EXTENT="31264"
BETYPE="BYTE" EXTTYPE="BYTE"/>
</fptr>

<div TYPE="DEPENDENT_WEB_RESOURCE">
```

```
<fptr
FILEID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-
update_www.gravatar.com_avatar_ad516503a11cd5ca435acc9bb6523536
_s_25_forcedefault_1_d_identicon">
    <area
FILEID="http___blogforever.eu_blog_gowers_2012_05_31_epsrc-
update-
update_www.gravatar.com_avatar_ad516503a11cd5ca435acc9bb6523536
_s_25_forcedefault_1_d_identicon" BEGIN="105" EXTENT="31264"
BETYPE="BYTE" EXTTYPE="BYTE"/>
</fptr>
</div>

<div TYPE="DEPENDENT_WEB_PAGE">
<fptr FILEID="http___blogforever.eu_blog_gowers_2012_04_13_a-
brief-epsrc-update_">
    <area
FILEID="http___blogforever.eu_blog_gowers_2012_04_13_a-brief-
epsrc-update_" BEGIN="105" EXTENT="31264" BETYPE="BYTE"
EXTTYPE="BYTE"/>
</fptr>
</div>

<div TYPE="DEPENDENT_WEB_PAGE">
<fptr
FILEID="http___www.epsrc.ac.uk_funding_fellows_Pages_areas.aspx
">
    <area
FILEID="http___www.epsrc.ac.uk_funding_fellows_Pages_areas.aspx
" BEGIN="105" EXTENT="31264" BETYPE="BYTE" EXTTYPE="BYTE"/>
</fptr>
</div>

<div TYPE="DEPENDENT_WEB_RESOURCE">
<fptr FILEID="http___s.stats.wordpress.com_w.js">
    <area FILEID="http___s.stats.wordpress.com_w.js"
BEGIN="105" EXTENT="31264" BETYPE="BYTE" EXTTYPE="BYTE"/>
</fptr>
</div>

<div TYPE="DEPENDENT_WEB_RESOURCE">
<fptr FILEID="http___blogforever.eu_blog_gowers_feed_">
    <area FILEID="http___blogforever.eu_blog_gowers_feed_"
BEGIN="105" EXTENT="31264" BETYPE="BYTE" EXTTYPE="BYTE"/>
</fptr>
</div>
</div>
</structMap>
<structLink>
  <smLink
xlink:from="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-update"
xlink:to="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-
```

```
update/www.gravatar.com/avatar/ad516503a11cd5ca435acc9bb6523536
?s=25&amp;forcedefault=1&amp;d=identicon"/>
  <smLink
xlink:from="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-update"
xlink:to="http://blogforever.eu/blog/gowers/2012/04/13/a-brief-
epsrc-update/"/>
  <smLink
xlink:from="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-update"
xlink:to="http://www.epsrc.ac.uk/funding/fellows/Pages/areas.as
px"/>
  <smLink
xlink:from="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-update" xlink:to="http://s.stats.wordpress.com/w.js"/>
  <smLink
xlink:from="http://blogforever.eu/blog/gowers/2012/05/31/epsrc-
update-update"
xlink:to="http://blogforever.eu/blog/gowers/feed/"/>

</structLink>
</mets>
<!-- </Appendix> -->
```

C. PREMIS in METS: an example

Numerous examples of schemas demonstrating use of PREMIS in METS are available at http://www.loc.gov/standards/premis/premis-mets.html.

In the ECHO Dep Generic METS Profile for Preservation and Digital Repository Interoperability , special attention has been given to administrative and technical metadata, particularly on integrating the PREMIS data model and schema into METS.

For example, the following string of code from that profile describes a single JPEG in PREMIS terms:

```
<techMD
        ID="APP1_TMD1PREMIS">
      – <mdWrap
          MDTYPE="PREMIS">
        – <xmlData>
          – <object
              xsi:schemaLocation="http://www.loc.gov/standards/premis/v1
http://www.loc.gov/standards/premis/v1/PR...">
            – <objectIdentifier>
                <objectIdentifierType>ECHODEP</objectIdentifierType>
                <objectIdentifierValue>BXF22.JPG</objectIdentifierValue>
              </objectIdentifier>
              <objectCategory>FILE</objectCategory>
            – <objectCharacteristics>
                <compositionLevel>0</compositionLevel>
              – <fixity>
                  <messageDigestAlgorithm>SHA-1</messageDigestAlgorithm>
                  <messageDigest>4638bc65c5b9715557d09ad373eefd147382ecbf</messageDigest>
                </fixity>
                <size>184302</size>
              – <format>
                – <formatDesignation>
                    <formatName>image/jpeg</formatName>
                    <formatVersion>1.02</formatVersion>
                  </formatDesignation>
                </format>
              </objectCharacteristics>
            </object>
          </xmlData>
        </mdWrap>
      </techMD>
```

D. Rights metadata in METS: an example

PREMIS Rights metadata should be used in the “rightsMD” METS section. If using all PREMIS units together the entire package goes in digiProvMD with the <premis> element as a container. An example on how PREMIS can be used in METS is shown in Table 1.

```
<mets:amdSec>
        <mets:rightsMD ID="ADM">
                <mets:mdWrap MDTYPE="OTHER" OTHERMDTYPE="PREMIS">
                <mets:xmlData>
                        <pre:rightsStatement>
                                <pre:rightsBasis>Copyright</pre:rightsBasis>
                                <pre:copyrightInformation>
                                        <pre:copyrightStatus>Under
                                copyright</pre:copyrightStatus>
                                        <pre:copyrightJurisdiction>us</pre:copyrightJurisdiction
                                >
                                        <pre:copyrightNote>Rights Holder(s):  Blogger Rachel
                                Beth Egenhoefer</pre:copyrightNote>
                                        <pre:copyrightNote> you’re more than welcome to
                                        steal it and repurpose it for your own use, just make
                                        sure to replace references to us with ones to you, and
                                        if you want we’d appreciate a link to Automattic.com
                                        somewhere on your site </pre:copyrightNote>

                                </pre:copyrightInformation>
                        </pre:rightsStatement>
                </mets:xmlData>
                </mets:mdWrap>
        </mets:rightsMD>
        ...
</mets:amdSec>
```

E. XML structure of the data model

The descriptions presented in this appendix represents the generic description of each of the record types that will be supported within the BlogForever repository. It is an intermediate generic description in preparation for inclusion into a METS object (see Appendix A).

E.1 Blog

```
<blog>
 <title desc="Title of the blog"/>
 <subtitle desc="Subtitles of the blog"/>
 <URI desc="URI of the blog"/>
 <status_code desc="Status defines whether the blog ceased to exist"/>
 <language desc="Retrieved language field, as defined by the blog"/>
 <charset desc="Retrieved character set field, as defined by the blog"/>
 <sitemap_uri desc="URI of the blog sitemap if exists"/>
 <platform desc="Platform of the blog powering service, retrieved where available"/>
 <platform_version desc="Versioning information about the platform"/>
 <webmaster desc="Information about the webmaster where available"/>
 <hosting_ip desc="IP address of the blog"/>
 <location_city desc="Location city based on the hosting details"/>
 <location_country desc="Location country based on the hosting details"/>
 <last_activity_date desc="Date as retrieved from the blog "/>
 <post_frequency desc="As retrieved from the blog"/>
 <update_frequency desc="As retrieved from the blog"/>
 <copyright desc="Notes of copyright as retrieved from the blog"/>
 <ownership_rights desc="Notes of ownership rights as retrieved from the blog"/>
 <distribution_rights desc="Notes of distribution rights as retrieved from the blog"/>
 <access_rights desc="Notes of access rights as retrieved from the blog"/>
 <blog_type desc="Type of the blog as defined by the selected blog
taxonomy"></blog_type>

 <datetime>
    <created></created>
    <first_captured></first_captured>
    <updated></updated>
    <timezone></timezone>
    <format></format>
 </datetime>
</blog>
```

E.2 Blog post

```
<post>
    <title desc="Title of the entry"></title>
    <subtitle desc="Subtitle of the entry if available"></subtitle>
    <URI desc="Entry URI"></URI>
    <date_created desc="Retrieved from the blog or obtained from the date or time
crawling"></date_created>
    <date_modified desc="Retrieved from the blog or obtained from the date/time
crawling"></date_modified>
    <version desc="Auto-increment: derived version number (versioning
support)"></version>
```

```
    <status_code desc="Information about the state of the post: active, deleted, updated
(versioning support)"></status_code>
    <geo_longitude desc="Geographic positioning information "></geo_longitude>
    <geo_latitude desc="Geographic positioning information"></geo_latitude>
    <visibility desc="Information about accessibility of the post"></visibility>
    <has_reply desc="Derived property (also SIOC )"></has_reply>
    <last_reply_date desc="Derived property (also SIOC)"></last_reply_date>
    <num_of_replies desc="Derived property (also SIOC)"></num_of_replies>
    <child_of desc="ID of entry parent if available "></child_of>
    <UI desc="Unique identification number assigned for enabling referencing"></UI>

    <type desc="Type of the post if specified (e.g. WordPress): attachment, page/post or
other custom type"></type>
    <posted_via desc="Information about the service used for posting if
specified"></posted_via>
    <previous_URI desc="URI to the previous post is available"></previous_URI>
    <next_URI desc="URI to the next post if available"></next_URI>

    <datetime>
      <created></created>
      <first_captured></first_captured>
      <updated></updated>
      <timezone></timezone>
      <format></format>
    </datetime>

    <author_list>
     <author>
      <name_displayed desc="Name of the poster as displayed"></name_displayed>
      <email_displayed desc="Email address of the poster as displayed
"></email_displayed>
      <is_anonymous desc="Boolean property to indicate anonymity"></is_anonymous>

      <community>
       Defined Separately
      </community>
     </author>
    </author_list>
</post>
```

E.3 Comment

```
    <comment>
      <type desc="Comment Type defined by a selected taxonomy of comments"></type>
      <content>See Entry for details</content>
      <external_comment_source_URI desc="URI for the source of the comment if
external"></external_comment_source_URI>
      <external_comment_service_name desc="Name of the service for the external
comment"></external_comment_service_name>

      <datetime>
        <created></created>
        <first_captured></first_captured>
        <updated></updated>
```

```
            <timezone></timezone>
            <format></format>
        </datetime>

        <author_list>
          <author>
            <name_displayed desc="Name of the poster as displayed"></name_displayed>
            <email_displayed desc="Email address of the poster as displayed "></email_displayed>
            <is_anonymous desc="Boolean property to indicate anonymity"></is_anonymous>

            <community>
            </community>
          </author>
        </author_list>
      </comment>
```

E.4 Page

```
  <page>
    <title desc="Title of the entry"></title>
    <subtitle desc="Subtitle of the entry if available"></subtitle>
    <URI desc="Entry URI"></URI>
    <date_created desc="Retrieved from the blog or obtained from the date/time crawling"></date_created>
    <date_modified desc="Retrieved from the blog or obtained from the date/time crawling"></date_modified>
    <version desc="Auto-increment: derived version number (versioning support)"></version>
    <status_code desc="Information about the state of the post: active, deleted, updated (versioning support)"></status_code>
    <geo_longitude desc="Geographic positioning information "></geo_longitude>
    <geo_latitude desc="Geographic positioning information"></geo_latitude>
    <visibility desc="Information about accessibility of the post"></visibility>
    <has_reply desc="Derived property (also SIOC )"></has_reply>
    <last_reply_date desc="Derived property (also SIOC)"></last_reply_date>
    <num_of_replies desc="Derived property (also SIOC)"></num_of_replies>
    <child_of desc="ID of entry parent if available "></child_of>

    <template desc="Information about the design template if available and if different from the general blog"></template>

    <author_list>
      <author>
        <name_displayed desc="Name of the poster as displayed"></name_displayed>
        <email_displayed desc="Email address of the poster as displayed "></email_displayed>
        <is_anonymous desc="Boolean property to indicate anonymity"></is_anonymous>

        <community>
        </community>
      </author>
    </author_list>
  </page>
```

E.5 Categorised Content

```
<categorised_content desc="Primary content, categorised by type" >
   <uri desc="URI of resource"></uri>
   <title desc="Title of the resource"></title>
   <is_embedded desc="Boolean value to indicate whether the resource is embedded"></is_embedded>
   <description desc="Description of the resource acquired from the crawled data"></description>
   <geo_latitude desc="Associated GEO positioning information where available"></geo_latitude>
   <geo_longitude desc="Associated GEO positioning information where available"></geo_longitude>
   <creator desc="Information about the creator where available"></creator>
   <file_path desc="File path to the media as stored on the disk"></file_path>
   <restriction desc="Requires extension to specify age, country or technical restrictions"></restriction>

   <rights>
     <copyright desc="Notes of copyright as retrieved from the blog"></copyright>
     <ownership_rights desc="Notes of ownership rights as retrieved from the blog"></ownership_rights>
     <distribution_rights desc="Notes of distribution rights as retrieved from the blog"></distribution_rights>
     <access_rights desc="Notes of access rights as retrieved from the blog"></access_rights>
     <licence desc="Licence of the content"></licence>
   </rights>

</categorised_content>
```

E.5.1 Image

```
  <image>
      <format desc="Image format"></format>
      <humbnail_uri desc="URI of the thumbnail associated with the acquired image"></humbnail_uri>
       <thumbnail_path desc="File path to the thumbnail of an image as stored on the disk"></thumbnail_path>
       <height desc="Dimensions of the image"></height>
       <width desc="Dimensions of the image"></width>
       <additional_meta_info desc="Additional columns to capture the necessary metadata for images as found necessary"></additional_meta_info>
     </image>
```

E.5.2 Video

```
   <video>
      <codec desc="Information about the codec of the video"></codec>
      <format desc="Format of the video file"></format>
      <duration desc="Duration of the video"></duration>
      <thumbnail_uri desc="URI of the thumbnail image for the video"></thumbnail_uri>
```

```
    <thumbnail_path desc="File path to the thumbnail of an image as stored on the disk"></thumbnail_path>
    <resolution desc="Information about the resolution of the video"></resolution>
    <additional_meta_i desc="Additional columns to capture the necessary metadata for images as found necessary"></additional_meta_i>
  </video>
```

E.5.3 Document

```
<document>
    <format desc="Format of the document file"></format>
    <language desc="Language in which the document is written or candidate"></language>
    <abstract desc="Abstract of the document or excerpt"></abstract>
    <text desc="The content of the document"></text>
</document>
```

E.5.4 Audio

```
<audio>
    <format desc="File format of the audio"></format>
    <bit_rate desc="Bit rate of the audio"></bit_rate>
    <duration desc="Duration of the audio track"></duration>
    <additional_meta_info desc="Additional columns to capture the necessary metadata for images as found necessary "></additional_meta_info>
</audio>
```

E.5.5 Tags

```
<tag>
  <tag desc="Tag that was added by a user"></tag>
  <language desc="Language of the tag"></language>
</tag>
```

E.5.6 Links

```
<link>
  <title desc="Title of the link if available"></title>
  <type desc="Recognized link types as identified from the data"></type>
  <URI desc="The value of the link"></URI>
  <rel desc="Recognised link relationship between resources"></rel>
  <rev desc="Reverse link relationship between resources"></rev>
</link>
```

E.5.7 Text

```
<text>
  <format desc="Information on text formatting as extracted from documents"></format>
```

```
    <language desc="Language in which the text is written"></language>
    <abstract desc="Abstract or excerpt from the text if available"></abstract>
    <text desc="Textual content"></text>
    <rights>
      <copyright desc="Notes of copyright as retrieved from the blog"></copyright>
      <ownership_rights desc="Notes of ownership rights as retrieved from the
blog"></ownership_rights>
      <distribution_rights desc="Notes of distribution rights as retrieved from the
blog"></distribution_rights>
      <access_rights desc="Notes of access rights as retrieved from the
blog"></access_rights>
      <licence desc="Licence of the content"></licence>
    </rights>
  </text>
```

E.5.8 Event

```
  <event>
    <name desc="Name of the event as identified form the crawled data"></name>
    <location desc="Location of the event or compound address"></location>
    <event_uri desc="Main URI describing the event"></event_uri>
    <date desc="Date and time of the event"></date>
    <affiliation desc="Organisation, companies, groups the event is affiliated"></affiliation>
    <type desc="Event type categorising the events or candidate entity"></type>
  </event>
```

F. List of Initiatives in Digital Preservation and Web Archiving

F.1 Projects and Collaborative Networks

APARSEN (http://www.alliancepermanentaccess.org/):

Network of Excellence gathering digital preservation practitioners and researchers. The research will include efforts to standardise authenticity protocols, and develop investigation into coordinated persistent identifiers.

ARCOMEM (http://www.arcomem.eu/):

Aiming to leverage the Wisdom of the Crowds for content appraisal, selection and preservation, in order to create and preserve archives that reflect collective memory and social content perception, and are, thus, closer to current and future users.

CAMiLEON (http://www2.si.umich.edu/CAMILEON/):

Conducting user studies and cost analysis for preservation strategies with respect to digital materials.

CASPAR (http://www.casparpreserves.eu)

Objectives, for example, include aims to: implement, extend, and validate the OAIS reference model; develop tools for capturing preservation related information; build visualisation services to support preservation; integrate rights management, authentication, and accreditation as features of OAIS; contribute to standardisation activities; Raise awareness of the need for digital preservation within the user-community;

Cedars (http://www.ukoln.ac.uk/services/elib/projects/cedars/):

Investigating digital information resources included in library collections that support preserving holdings over the long-term.

Digital Curation Centre (http://www.dcc.ac.uk):

The initiative is funded by the Joint Information Systems Committee (http://www.jisc.ac.uk) and is a “centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management” .

Digital Preservation Coalition (http://www.dpconline.org/):

Offers membership to institutes engaged in areas of digital preservation to promote worforce development and capacity building, encourage knowledge exchange, develop assurance and practice, and build partnership and sustainability.

International Internet Preservation Consortium (http://netpreserve.org/):

Aims to “acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere” .

InSPECT (http://www.significantproperties.org.uk/):

Identifying the functions performed by an Object in its current manifestation and evaluating if they are required by other stakeholders.

InterPARES (http://www.interpares.org/)

Addressing problems regarding the preservation of the authenticity of electronic records that are: no longer needed by the creating body to fulfill its own mandate; in the context of artistic, scientific and government activities that are conducted using experiential, interactive and dynamic computer technology; in digital systems in small and medium-sized archival organizations.

KEEP (http://www.keep-project.eu/):

Wordcloud: could not retrieve content from this or any of the following pages: http://www.keep-project.eu/ezpub2/index.php, http://www.keep-project.eu/ezpub2/index.php?/eng/About-KEEP, http://www.keep-project.eu/ezpub2/index.php?/eng/Work-Packages

Creating portable emulators enabling access to and use of digital objects stored on outdated computer media. The emulators will ensure accurate rendering of both static and dynamic digital objects.

LAWA (http://www.lawa-project.eu/):

Aims to develop sustainable infra-structure and usable software tools for aggregating, querying, and analyzing data on the Internet.

Linked Data (http://linkeddata.org/):

Introduces a workflow for exposing and sharing information and knowledge using URIs and RDF.

LiWA (http://liwa-project.eu/):

Developing and demonstrating web archiving tools able to capture content from a wide variety of sources, to improve archive fidelity and authenticity and to ensure long term interpretability of web content.

LOCKSS (http://lockss.org):

The LOCKSS Program is a library-led digital preservation system built on the principle that “lots of copies keep stuff safe.”

Memento (http://mementoweb.org/)

Making it straightforward to access the Web of the past as it is to access the current Web. Creating a framework to link resources in a page to existing ones on the web around selected periods.

Open Planets Foundation (http://www.openplanetsfoundation.org/):

Providing practical solutions and expertise in digital preservation, building on the research and development outputs of the Planets project.

PARADIGM (http://www.paradigm.ac.uk/):

Building templates, testing tools, and setting up best practices for the long-term preservation of material in a personal digital archive.

Parse.insight (http://www.parse-insight.eu/)

Concerned with the preservation of digital information in science, from the preservation of raw data through to the final publications resulting from a study of the data.

Papyrus (http://www.ict-papyrus.eu/):

Exploring issues related to interoperability and preservation, where one might draw content from one domain to make it available for users in another domain.

PersID (http://www.persid.org/)

Investigating ways to assign unambiguous persistent identifiers to scholarly and cultural information.

Planets (http://www.planets-project.eu/):

The planets project was founded to create explicit workflows to aid decision-making about long term preservation, and encourage increased automation and introduce scalable infrastructure.

PrestoPRIME (http://www.prestoprime.org/):

Addressing long-term preservation of and access to digital audiovisual content by integrating media archives with European online digital libraries. The research resulted in a range of tools and services, delivered through the networked Competence Centre PrestoCentre.

PrestoSpace (http://www.prestospace.org/)

Working toward producing sustainable assets with easy access for larger exploitation and distribution to specialists and general public, driven by the idea that an accessible item is more valuable than an item on a shelf and more likely to be maintained.

PROTAGE (http://www.protage.eu/):

Building and validating software agents for long-term digital preservation and access that can be integrated in existing and new preservation systems. Investigating digital objects independent of software and hardware technology. Intelligent objects.

SCAPE (http://www.scape-project.eu/):

Enhance the state of the art of digital preservation in three ways: by developing an infrastructure and tools for scalable preservation actions; by providing a framework for automated, quality-assured preservation workflows and by integrating these components with a policy-based preservation planning and watch system.

SHAMAN (http://shaman-ip.eu/):

Developing a next generation digital preservation framework including tools for analysing, ingesting, managing, accessing and reusing information objects and data across libraries and archives.

SCIDIP-ES (http://www.scidip-es.eu)

Creating infrastructure for e-science that includes science data preservation.

SPRUCE (http://wiki.opf-labs.org/display/SPR/Home, http://www.dpconline.org/advocacy/spruce):

Organising agile development mashups providing technical support for real digital preservation challenges that institutions face.

TIMBUS (http://timbusproject.net/):

Ensuring continued access to services and software necessary to produce the context within which information can be accessed, properly rendered, validated and transformed into knowledge.

Wf4Ever (http://www.wf4ever-project.org/):

Providing the methods and tools required to ensure the long-term preservation of scientific workflows.

F.2 Web Archives: National and Event

Comprehensive list at:

http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

from this list:

National level

Library of Congress Minerva Collection
(http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html) p
Australia (http://pandora.nla.gov.au/)
New Zealand (http://www.natlib.govt.nz/collections/a-z-of-all-collections/nz-web-archive) p
Austria (http://www.onb.ac.at/ev/about/webarchive.htm)
Canada (http://www.collectionscanada.gc.ca/index-e.html) empty
Croatia (http://haw.nsk.hr/) p
Czech Republic (http://en.webarchiv.cz/)
Denmark (http://netarkivet.dk/)
Finland (http://verkkoarkisto.kansalliskirjasto.fi/)
France (http://www.bnf.fr/en/professionals/digital_legal_deposit.html) p
German Bundestag (http://webarchiv.bundestag.de/) tp
Iceland (http://vefsafn.is/)
Netherlands (http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/)
Latvia (http://www.lnb.lv/lv/par-lnb/struktura/bibliografijas-instituts)
Portugal (http://www.archive.pt/)
Căcăk, Serbia (http://digital.cacak-dis.rs/english/web-archive-of-cacak/)
Slovenia (http://www.zal-lj.si/)
Spain (http://www.bne.es/es/LaBNE/PreservacionDominioES/) have to check
Catalonia (http://www.padicat.cat/) p
Sweden (http://www.kb.se/english/find/internet/websites/)
Switzerland (https://www.e-helvetica.nb.admin.ch/)
UK (http://www.webarchive.org.uk/ukwa/)
UK Government Website Archive (http://www.nationalarchives.gov.uk/webarchive/)
Library of Congress (http://www.loc.gov/webarchiving/)
Greece (http://archive.aueb.gr/)
Russia (http://www.opengovdata.ru/archive/)

Other

North Carolina State Government
(http://www.archive-it.org/collections/north_carolina_state_government_web_site_archive)
Virginia State Judicial Branch
(http://www.archive-it.org/collections/virginia_state_government_judicial_branch_collection)
Internet Memory (http://internetmemory.org/en/) p
California Digital Library (http://webarchives.cdlib.org/)
Internet Archive (http://archive.org/)
Columbia University
(https://www1.columbia.edu/sec/cu/libraries/bts/web_resource_collection/index.html)
Harvard University
(http://wax.lib.harvard.edu/collections/home.do) p
University of Michigan (http://bentley.umich.edu/uarphome/webarchives/webarchive.php)
University of Texas at San Antonio
(http://www.archive-it.org/public/partner.html?id=318)
World Bank (http://go.worldbank.org/67KZ5AH4Y0)
Hurricane Katrina & Rita (http://websearch.archive.org/katrina/)
2004 Presidential Term Web Harvest (http://web.resourceshelf.com/go/resourceblog/43866)
Anarchism (http://www.archive-it.org/collections/anarchism)
Archive of Venezuelan Political Discourse
(http://www.archive-it.org/collections/archive_of_venezuelan_political_discourse_arvepodis)
University of Southern California
(http://www.archive-it.org/collections/university_of_southern_california_website_archive)
University of Toronto
(http://www.archive-it.org/collections/university_of_toronto_web_archives)
Canadian Labour Unions
(http://www.archive-it.org/collections/canadian_labour_unions)
Islamic Middle East (http://www.archive-it.org/collections/islamic_middle_east)
Latin American Government Documents Archive
(http://www.archive-it.org/collections/latin_american_government_documents_archive_lagda)
Clinton Library White House Website “Snap Shots” (http://128.83.78.246/archivesearch.html)

Collections aggregated using Archive-It:

http://www.archive-it.org/explore/?show=Collections