Blogs as Objects of Preservation: Advancing the Discussion on Significant Properties

Karen Stepanyan
Department of Computer Science, University of Warwick, Coventry, CV47AL, UK
K.Stepanyan@warwick.ac.uk

George Gkotsis
Department of Computer Science, University of Warwick, Coventry, CV47AL, UK
G.Gkotsis@warwick.ac.uk

Hendrik Kalb
Institute for Business Informatics, Technische Universität Berlin, Germany
Hendrik.Kalb@ tu-berlin.de

Yunhyong Kim
School of Humanities, University of Glasgow, UK
Yunhyong.Kim@glasgow.ac.uk

Alexandra I. Cristea
Department of Computer Science, University of Warwick, Coventry, CV47AL, UK
A.I.Cristea@warwick.ac.uk

Mike Joy
Department of Computer Science, University of Warwick, Coventry, CV47AL, UK
M.S.Joy@warwick.ac.uk

Matthias Trier
Department of IT Management, Copenhagen Business School, Denmark
mt.itm@cbs.dk

Seamus Ross
Faculty of Information, University of Toronto, Canada
Seamus.Ross@utoronto.ca

ABSTRACT
The quest for identifying 'significant properties' is a common challenge for 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. This paper advances earlier work by building on the existing InSPECT framework and improving its capabilities of working with complex/compound objects like blogs. The modifications enable a more thorough analysis of object structures, accentuate the differences and similarities between the framework's two streams of analysis (i.e. Object and Stakeholder analysis) and, subsequently, improve the final reformulation of the properties. To demonstrate the applicability of the modified framework, the paper presents a use case of a blog preservation initiative that is informed by stakeholder interviews and evaluation of structural and technological foundations of blogs. It concludes by discussing the limitations of the approach and suggesting directions for future research.

Categories and Subject Descriptors
H.3.7 Digital Libraries

General Terms
Design, Theory

Keywords
Blogs, Weblogs, Digital Preservation, Significant Properties

1. INTRODUCTION
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 [1]. 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.

The ephemeral nature of web resources encouraged the development of long-term accessibility and preservation actions such as the Internet Archive[1] or HTTP Archive[2]. Web archiving initiatives, such as Arcomem[3] or LiWA[4], 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. This paper 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 paper proceeds to highlight the limitations within existing research on identifying these properties and proposes improvements accordingly. The paper concludes by demonstrating the application of the modified approach on a use case and discussing the benefits and limitations of the proposed approach.

[1] http://archive.org
[2] http://httparchive.org/
[3] http://www.arcomem.eu/
[4] http://liwa-project.eu/

2. RELATED WORK

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 [2]. Another example that justifies preservation initiatives is the account of disappearing personal diaries. Their loss is believed to have implications for our cultural memory [3]. 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.

Hank and her colleagues [4, 5] 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 [6, p. 43]. Other definitions, for instance by Pluempavarn and Panteli [7, 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 [8] 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?

2.1 What to preserve in blogs: significant properties

In the digital preservation community, one of the prevailing approaches for defining what to preserve is bound to the notion of significant properties[5] [9]. It is argued [10] that significant properties can help define the object and specify what to preserve, before deciding how to preserve. It has been acknowledged [11], 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, Low [12] 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[6] Plato that focuses on stakeholder requirements [13], InSPECT that combines object and stakeholder analysis [14], a JISC[7]-funded initiative that continues the discussion [15], and a template of characteristics [16] developed by NARA[8].

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 [17].

However, considering the range of available solutions, InSPECT framework [14] is considered to offer a more balanced approach to identifying significant properties [12]. 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.

2.2 Limitations of the Base Framework

InSPECT [14] is built on the Function-Behaviour-Structure framework (FBS) [18] 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 [19], 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 behaviors". Similar complexities exist in the context of blogs, 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

[5] http://www.leeds.ac.uk/cedars/

[6] http://www.planets-project.eu/

[7] www.jisc.ac.uk/

[8] http://www.archives.gov/

use of characterisation tools or technical specifications for the purpose of object structure analysis (Section 3.1 of [12]). 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 markup, 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 behavior 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. PROPOSED CHANGES TO PRESERVATION PERSPECTIVES

The modifications discussed in this paper, 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 paper. 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.

3.1 Benefits of Ontological Perspectives

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 [20] 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.

3.2 Description of the Framework

This section outlines each stage of the workflow and describes the major streams of analysis in greater detail. We focus on the modified parts of the framework, referring the reader to documentation of InSPECT for further details.

The diagrammatic representation of the proposed framework is presented in Fig. 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 [21] 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.

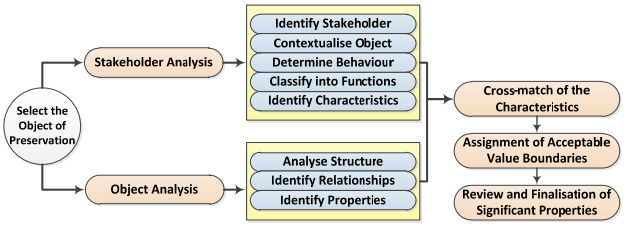

Fig. 1: Modified version of the base framework.

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).

4. Use Case: Blog Preservation

This section integrates and consolidates some of the work carried forward as part of a blog preservation initiative [22, 23]. 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.

4.1 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) [24]. 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 [25]. 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 from this paper. We briefly outline the conducted work, the details of which are available elsewhere [see 23].

4.1.1 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.

4.1.2 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.

Within the context of this paper, an evaluation of 209,830 blog pages has been performed [26]. 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 [23] due to space constraints.

4.2 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 is available elsewhere [22]. A brief outline of the process directly affecting this use case is presented below.

4.2.1 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 [27]. 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.

4.2.2 Applied Method of Requirement Analysis.

There is a range of methods for conducting effective user requirement analysis [28]. 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.

4.2.3 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 [22]. The requirements that depend on existence of certain data elements were then shortlisted as shown in Table 1.

Table 1: 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 2. 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 2: 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.*

4.3 Cross-Matching and Refining Properties.

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/UI9 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.

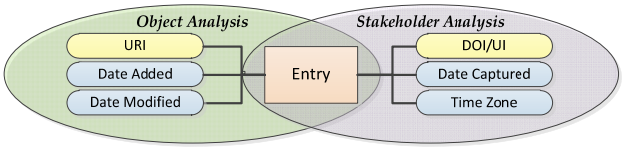

Fig. 2: An example of cross-matching and integration of properties, which were identified from the two parallel streams of Object and Stakeholder Analysis.

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 2.

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

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. 2.

4.3.1 Review and Finalisation of Properties

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.

5. DISCUSSION

The use case (Section 4) 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 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.

6. CONCLUSIONS

This paper advances 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 paper 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 paper adapted the recently developed InSPECT framework [12] 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 emerging from the study summarised in this paper 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.

Acknowledgments: *This work was conducted as part of the BlogForever project co-funded by the European Commission Framework Programme 7 (FP7), grant agreement No.269963.*

7. REFERENCES

[1] Garden, M. Defining blog: A fool's errand or a necessary undertaking. *Journalism*(20 September 2011 2011), 1-17.
[2] Chen, X. Blog Archiving Issues: A Look at Blogs on Major Events and Popular Blogs. *Internet Reference Services Quarterly*, 15, 1 2010), 21-33.
[3] 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.
[4] Sheble, L., Choemprayong, S. and Hank, C. *Surveying bloggers' perspectives on digital preservation: Methodological issues*. City, 2007.
[5] Hank, C. *Blogger perspectives on digital preservation: Attributes, behaviors, and preferences*. City, 2009.
[6] Nardi, B. A., Schiano, D. J., Gumbrecht, M. and Swartz, L. Why we blog. *Communications of the ACM*, 47, 12 2004), 41-46.
[7] Pluempavarn, P. and Panteli, N. *Building social identity through blogging*. Palgrave Macmillan, City, 2008.
[8] Lomborg, S. Navigating the blogosphere: Towards a genre-based typology of weblogs. *First Monday*, 14, 5 2009).
[9] 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.
[10] Deken, J. M. Preserving Digital Libraries. *Science & Technology Libraries*, 25, 1-2 (2004/11/29 2004), 227-241.
[11] Knight, G. and Pennock, M. Data without meaning: Establishing the significant properties of digital research. *International Journal of Digital Curation*, 4, 1 2009), 159-174.
[12] 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.
[13] Becker, C., Kulovits, H., Rauber, A. and Hofman, H. *Plato: a service oriented decision support system for preservation planning*. ACM, City, 2008.
[14] Knight, G. *InSPECT framework report*. 2009.
[15] Hockx-Yu, H. and Knight, G. What to preserve?: significant properties of digital objects. *International Journal of Digital Curation*, 3, 1 2008), 141-153.
[16] NARA. *Significant Properties*. NARA, 2009.
[17] Farquhar, A. and Hockx-Yu, H. Planets: Integrated services for digital preservation. *International Journal of Digital Curation*, 21, 2 2007), 88-99.
[18] Gero, J. S. Design prototypes: a knowledge representation schema for design. *AI magazine*, 11, 4 1990), 26.
[19] Sacchi, S. and McDonough, J. P. *Significant properties of complex digital artifacts: open issues from a video game case study*. ACM, City, 2012.
[20] Doerr, M. and Tzitzikas, Y. Information Carriers and Identification of Information Objects: An Ontological Approach. *Arxiv preprint arXiv:1201.0385*2012).
[21] Dappert, A. and Farquhar, A. Significance is in the eye of the stakeholder. *Research and Advanced Technology for Digital Libraries*2009), 297-308.
[22] 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.
[23] Stepanyan, K., Joy, M., Cristea, A., Kim, Y., Pinsent, E. and Kopidaki, S. *D2.2 Report: BlogForever Data Model*. 2011.
[24] Thibodeau, K. *Overview of technological approaches to digital preservation and challenges in coming years*. Council on Library and Information Resources, City, 2002.
[25] 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.
[26] Banos, V., Stepanyan, K., Joy, M., Cristea, A. I. and Manolopoulos, Y. *Technological foundations of the current Blogosphere*. City, 2012.
[27] Yeo, G. 'Nothing is the same as something else': significant properties and notions of identity and originality. *Archival Science*, 10, 2 2010), 85-116.
[28] Hull, E., Jackson, K. and Dick, J. *Requirements engineering*. Springer-Verlag New York Inc, 2010.