

Establishing Interoperability of a Blog Archive through Linked Open Data

Hendrik Kalb¹, Paraskevi Lazaridou², Matthias Trier³

^{1,2}Fachgebiet Systemanalyse und EDV / ³Department of IT Management

^{1,2}Technische Universität Berlin / ³Copenhagen Business School

^{1,2}Einsteinufer 17 / ³Solbjerg Plads 3

^{1,2}10587 Berlin / ³2000 Frederiksberg

¹hendrik.kalb@tu-berlin.de

²paraskevi.lazaridou@tu-berlin.de

³mt.itm@cbs.dk

Abstract: The digital cultural heritage is partly preserved through web archiving activities. The BlogForever platform is a web archiving platform that aims specifically at the preservation of the blogosphere. The focus enables exploitation of the blog structure for sophisticated access capabilities on archived data. However, interoperability among BlogForever archives, as well as with other digital libraries, is necessary in order to avoid silos of data. In this paper, we reveal some of our efforts to establish interoperability through the application of Linked Open Data.

1 Introduction

The aim of this paper is to present how the BlogForever platform, or any other blog preservation platform, can establish interoperability through Linked Open Data.

Web archiving is an important aspect in the preservation of cultural heritage [Mas06] and, therefore, several projects from national and international organisations are working on web preservation activities. The blogosphere, as part of the web, has an increasing societal impact next to traditional media like press or TV. Prominent examples are the influential blogs in political movements in Egypt [Ish08] or Iran [Col05]. But there are also other domains that people engage in blogging, e.g. in the fields of arts or science [WJM10], or leisure activities [Chi10]. However, blogs, as other social media, are ephemeral and that becomes particularly obvious by the loss of blogs about major historical events in the recent past [Che10, Ent04]. Furthermore, the loss of personal diaries in the form of blogs has implications for our cultural memory [O'S05].

The BlogForever¹ project creates a novel software platform capable of aggregating, preserving, managing and disseminating blogs. Through the specialisation in blog archiving, as a subcategory of web archiving, the specific features of blogs can be exploited in order

¹<http://blogforever.eu>

to overcome limitations of current web archiving.

2 The BlogForever data model

While it seems that it is almost impossible to give an exclusive definition for the nature of blogs [Gar11, Lom09], it is necessary for preservation activities to identify blogs' properties [SGK⁺12]. This is even more crucial for the BlogForever platform which aims at sophisticated access capabilities for the archived blogosphere. Therefore, different appearances of blogs were examined, and a comprehensive data model was created.

The development of the data model was based on existing conceptual models of blogs, data models of open source blogging systems, an empirical study of web feeds, and an online survey on bloggers' and blog readers' perceptions. Thus, it was possible to identify various entities like core blog elements (e.g. blog, post, comments), embedded content (e.g. images, audio, video), links (e.g. embedded links, blogroll, pingback), layout (e.g. css, images), feeds (e.g. RSS, Atom), user profiles and affiliations [SJC⁺11].

The full model comprises over forty single entities and each entity is subsequently described by several properties, e.g. title, URI, aliases, etc. Figure 1 shows a high level view of the blog core. The directions of the relationships between the primary identified entities of a weblog are indicated with small triangles [SJC⁺11].

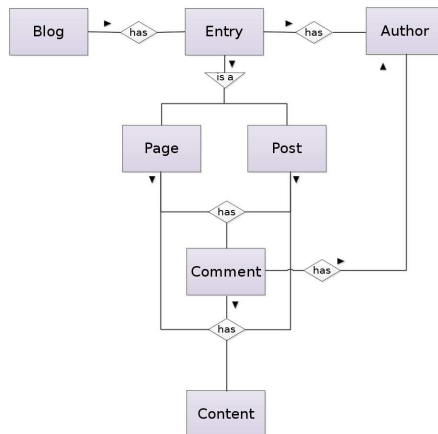


Figure 1: Core of the generic blog data model [SJC⁺11, p. 45]

Besides the inherent blog properties, additional metadata about archiving and preservation activities are captured, stored, and managed. For example, information regarding the time of harvesting of a blog or the legal rights of the content, have to be documented as well. Furthermore, additional data may emerge as well as annotations from the archive users, like tags or comments.

3 Interoperability scenario for the BlogForever platform

The successful completion of the BlogForever project anticipates the adoption of the archiving prototype by a diverse range of institutions that will deploy it to preserve collections of blogs. The selection depends on the specific preservation aims of each institution. Thus, interoperability between the collections will become an immediate concern in order to avoid silos of data that cannot be easily shared, re-organised, and/or re-used by end-users. The application of ontologies can enhance the interoperability by the provision of open standards for describing, accessing, and connecting data.

The interoperability of the BlogForever environment has to be considered on two levels. First, there should be interoperability among different BlogForever archives. For example, a retrieval process for weblog data could operate on several archives and the results of complex search queries can be merged automatically. Assume that a selected repository contains the blogs of the academic staff of a specified set of universities, and another repository preserves the blogs of the members of a selected scientific association. It is probable that the data of the two repositories overlap partially. The use of shared vocabularies and a common ontology would allow an application to automatically merge the data from both repositories, providing a user of the repository with the means of searching and exploring the data as if they are from the same repository.

Furthermore, interoperability with other external repositories, for example, digital libraries, could be also supported. Digital libraries contain endless amounts of data that can be related to the data preserved in a BlogForever archive. Unlike interoperability between two BlogForever archives, the connection with another digital library will extend the amount of concepts in the resulting ontology. In other words, two BlogForever archives share a common set of concepts (e.g. blog, post, blog author) and a merging means to merge instances of these concepts. However, a digital library has its own concepts like author, book, newspaper, etc. The relations between the concepts of both repositories have to be expressed (e.g. a blog author is a kind of author). Once the relations between the concepts are expressed formally, a merging of instances of both repositories will be possible.

4 Exposing BlogForever data as LOD

The interoperability scenario for the BlogForever platform can be established through exposing the BlogForever data as Linked Open Data² (LOD) [HB11]. Thereby, it is important that the exposure is not limited to descriptive metadata (e.g. title) but includes as well structural and administrative metadata [Cun04]. Several existing vocabularies have been considered for exposing data from a BlogForever repository as LOD. Dublin core³ (DC), friend of a friend⁴ (FOAF), and semantically-interlinked online communities⁵ (SIOC) have

²<http://linkeddata.org/>

³<http://dublincore.org/>

⁴<http://www.foaf-project.org/>

⁵<http://sioc-project.org/>

emerged as the most important. Additionally, the PREMIS⁶ vocabulary defines specific concepts for preserved objects.

While it is not possible here to reveal all the concepts that are used to expose the data based on the BlogForever data model as LOD, a subset of important classes and properties is presented. Table 1 shows exemplary some classes that are used to expose main entities in the core data model and important preservation activities. Table 2 presents a selection of properties for four types of metadata: descriptive, structural, administrative and provenance metadata..

Table 1: Classes for exposing BlogForever data

Class	Description	Vocabulary & URI
Weblog	The entire weblog. It must not be confused with individual posts or other parts of the blog.	http://rdfs.org/sioc/types#Weblog
BlogPost	A blog post. It must not be confused with the blog or comments in the blog.	http://rdfs.org/sioc/types#BlogPost
Comment	A comment on a blog post.	http://rdfs.org/sioc/types#Comment
Event	Information about an action on the resource, e.g. harvest, ingest, migration.	http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl#Event
Creating-Application	The original blog hosting platform, e.g. Wordpress.	http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl#CreatingApplication

5 Conclusion

In this paper, we considered the provision of data stored in BlogForever archives and their relationships as LOD with an explicit semantic that is machine readable. Thereby, several purposes will be served. First, different BlogForever archives would be able to interoperate easily on a data level. Second, they can be integrated with other digital libraries like databases for scientific publications to enable search queries and navigation that are not limited to the authors and publications in the blog repositories. Third, the data can be linked to other repositories in order to connect terms to public available definitions and descriptions of these terms. For example, the topic of a blog or blog post can be linked to the definition in Wikipedia or other repositories. Fourth, new relationships can be easily created and expressed (e.g. with SPARQL) based on the existing data and relationships. Fifth, the provision of the data in an openly available standardized format and with a unique identifier for each object facilitates the development of third party applications.

⁶<http://www.loc.gov/standards/premis/owlOntology-announcement.html>

Table 2: Properties for exposing BlogForever data

Property	Domain of the property	Description	Vocabulary & URI
Descriptive Metadata			
title	Weblog, BlogPost	The title of a blog or a blog post.	http://purl.org/dc/terms/title
creation date	Weblog, BlogPost, Comment	The date when the resource has been created.	http://purl.org/dc/terms/created
language	Weblog, BlogPost, Comment	The language of the resource.	http://purl.org/dc/terms/language
creator	BlogPost, Comment	The author of the resource.	http://purl.org/dc/terms/creator
type / genre	Weblog	Nature or genre of the blog.	http://purl.org/dc/terms/type
Structural Metadata			
container_of	Weblog	Blog posts or comments that the weblog contains. The inverse of has_container.	http://rdfs.org/sioc/ns#container_of
reply_of	Comment	The blog or comment to which the comment answers.	http://rdfs.org/sioc/spec/#reply_of
links_to	BlogPost, Comment	Resources that are linked in the blog post or comment.	http://rdfs.org/sioc/spec/#links_to
Administrative Metadata			
source	Weblog, BlogPost, Comment	The original URI where the resource has been crawled.	http://purl.org/dc/terms/source
accessRight	Weblog, BlogPost, Comment	Information about who can access the resource or an indication of its security status.	http://purl.org/dc/terms/accessRights
license	Weblog, BlogPost, Comment	A legal document giving official permission to do something with the resource.	http://purl.org/dc/terms/license
format	Weblog, BlogPost, Comment	The file format, physical medium, or dimensions of the resource.	http://purl.org/dc/terms/format
checksum_type	BlogPost, Comment	The algorithm used for check sum.	http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl#messageDigestAlgorithm
checksum_value	BlogPost, Comment	Hash sum value calculated using checksum algorithm.	http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl#messageDigest
Provenance Metadata			
creating-Application-Name	CreatingApplication	Name of the blog hosting platform, e.g. Wordpress.	http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl#creatingApplicationName
creating-Application-Version	CreatingApplication	Version of the blog hosting platform.	http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl#creatingApplicationVersion
eventType	Event	The nature of the event, e.g. harvest, ingest, migration.	http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl#eventType
eventDateTime	Event	The date and time when the event occurred.	http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis.owl#eventDateTime

6 Acknowledgments

This work was conducted as part of the BlogForever⁷ project co-funded by the European Commission Framework Programme 7 (FP7), grant agreement No.269963.

References

- [Che10] Xiaotian Chen. Blog Archiving Issues: A Look at Blogs on Major Events and Popular Blogs. *Internet Reference Services Quarterly*, 15(1):21–33, February 2010.
- [Chi10] Tara Chittenden. Digital dressing up: modelling female teen identity in the discursive spaces of the fashion blogosphere. *Journal of Youth Studies*, 13(4):505–520, August 2010.
- [Col05] Stephen Coleman. Blogs and the New Politics of Listening. *The Political Quarterly*, 76(2):272–280, April 2005.
- [Cun04] Morgan V Cundiff. An introduction to the Metadata Encoding and Transmission Standard (METS). *Library Hi Tech*, 22(1):52–64, 2004.
- [Ent04] Richard Entlich. Blog Today, Gone Tomorrow? Preservation of Weblogs. *RLG Dig-iNews*, 8(4), 2004.
- [Gar11] M Garden. Defining blog: A fool’s errand or a necessary undertaking. *Journalism*, September 2011.
- [HB11] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 1st edition, 2011.
- [Ish08] Tom Isherwood. A new direction or more of the same? Political blogging in Egypt. *Arab Media & Society*, September 2008.
- [Lom09] Stine Lomborg. Navigating the blogosphere: Towards a genre-based typology of weblogs. *First Monday*, 14(5), May 2009.
- [Mas06] Julien Masanès. *Web Archiving*. Springer-Verlag, Berli, Heidelberg, 2006.
- [O’S05] Catherine O’Sullivan. Diaries, On-line Diaries, and the Future Loss to Archives; or, Blogs and the Blogging Bloggers Who Blog Them. *The American Archivist*, 68(1):53–73, 2005.
- [SGK⁺12] Karen Stepanyan, George Gkotsis, Hendrik Kalb, Yunhyong Kim, Alexandra I. Cristea, Mike Joy, Matthias Trier, and S amus Ross. Blogs as Objects of Preservation: Advancing the Discussion on Significant Properties. In *iPres 2012*, pages 218–224, Toronto, Canada, 2012.
- [SJC⁺11] Karen Stepanyan, Mike Joy, Alexandra Cristea, Yunhyong Kim, Ed Pinsent, and Stella Kopidaki. Blogforever: Weblog Data Model. Technical report, 2011.
- [WJM10] Xiaoguang Wang, Tingting Jiang, and Feicheng Ma. Blog-supported scientific communication: An exploratory analysis based on social hyperlinks in a Chinese blog community. *Journal of Information Science*, 36(6):690–704, December 2010.

⁷<http://blogforever.eu/>