UROBE: A PROTOTYPE FOR WIKI PRESERVATION

Niko Popitsch, Robert Mosser, Wolfgang Philipp
University of Vienna
Faculty of Computer Science

ABSTRACT

More and more information that is considered for digital long-term preservation is generated by Web 2.0 applications like wikis, blogs or social networking tools. However, there is little support for the preservation of these data today. Currently they are preserved like regular Web sites without taking the flexible, lightweight and mostly graph-based data models of the underlying Web 2.0 applications into consideration. By this, valuable information about the relations within these data and about links to other data is lost. Furthermore, information about the internal structure of the data, e.g., expressed by wiki markup languages is not preserved entirely.

We argue that this currently neglected information is of high value in a long-term preservation strategy of Web 2.0 data and describe our approach for the preservation of wiki contents that is based on Semantic Web technologies. In particular we describe the distributed architecture of our wiki preservation prototype (*Urobe*) which implements a migration strategy for wiki contents and is based on Semantic Web and Linked Data principles. Further, we present a first vocabulary for the description of wiki core elements derived from a combination of established vocabularies/standards from the Semantic Web and digital preservation domains, namely Dublin Core, SIOC, VoiD and PREMIS.

1 INTRODUCTION

Users of Web 2.0 applications like wikis, blogs or social networking tools generate highly interlinked data of public, corporate and personal interest that are increasingly considered for long-term digital preservation. The term Web 2.0 can be regarded as referring to "a class of Web-based applications that were recognized *ex post facto* to share certain design patterns", like being *user-centered*, *collaborative* and *Web-based* [6]. Web 2.0 applications are usually based on flexible, lightweight data models that interlink their core elements (e.g., users, wiki articles or blog posts) using hyperlinks and expose these data on the Web for human consumption as HTML. The current practice for the preservation of these data is to treat this layer like a regular Web site and archive the HTML representations (usually by crawling them) rather than the core elements themselves. By this, some irrelevant information (like e.g., automatically generated pages) is archived while some valuable information about the semantics of relationships between these elements is lost or archived in a way that is not easily processable by machines [1] . For example, a wiki article is authored by many different users and the information who authored what and when is reflected in the (simple) data model of the wiki software. This information is required to access and integrate these data with other data sets in the future. However, archiving only the HTML version of a *history* page in Wikipedia makes it hard to extract this information automatically.

Another issue is that the internal structure of particular core elements (e.g., wiki articles) is currently not preserved adequately. Wiki articles are authored using a particular wiki markup language. These simple description languages contain explicit information about the structure of the text (e.g., headings, emphasized phrases, lists and tables, etc.). This internal structure is lost to some extent if digital preservation strategies consider only the HTML version of such articles rendered by a particular wiki software as this rendering step is not entirely reversible in many cases.

In a summary, we state that the current practice for the preservation of Web 2.0 data preserves only one particular (HTML) representation of the considered data instead of preserving the core elements of the respective data models themselves. However, we consider these core elements and their relations crucial for future data migration and integration tasks. In the following we introduce our system Urobe that is capable of preserving the core elements of data that are created using various wiki software.

2 UROBE: A WIKI PRESERVATION TOOL

We are currently developing a prototype (Urobe) for the long-term preservation of data created by wiki users. One particular problem when considering wiki preservation is that there exists not one single but a large number of different wiki implementations [2] , each using its own wiki markup language. This is what makes a general emulation approach for preserving wiki contents unfeasible as it would require establishing appropriate emulation environments for each wiki software. After further analyzing

[1] Cf. `http://jiscpowr.jiscinvolve.org/wp/2009/03/25/arch-wiki/`

[2] For example, the website `http://www.wikimatrix.org/` lists over 100 popular wiki engines.

several popular wiki engines, we have identified the following required components for implementing a long-term, migration-based wiki preservation strategy:

1. An abstract, semantic *vocabulary / schema* for the description of core elements and their relations stored in a wiki, namely: users, articles and revisions, their contents, links, and embedded media.
2. Software components able to *extract these data* from wiki implementations.
3. A scalable infrastructure for *harvesting and storing* these data.
4. Migration services for *migrating contents* expressed in a wiki markup language into standardized formats.
5. Migration services for the *semantic transformation* of the meta data stored in this system to newer formats (i.e., services for vocabulary evolution).
6. Software interfaces to existing digital preservation infrastructures using preservation meta data standards.
7. An effective user interface for accessing these data.

2.1 Benefits of Semantic Web Technologies

Urobe is implemented using Semantic Web technologies: It extracts data stored in a wiki and archives it in the form of named RDF graphs [4]. The resources and properties in these graphs are described using a simple OWL [3] vocabulary. Resources are highly interlinked with other resources due to structural relationships (e.g., article revisions are linked with the user that authored them) but also semantic relationships (e.g., user objects stemming from different wikis that are preserved by Urobe are automatically linked when they share the same e-mail address). We decided to make use of Semantic Web technologies for the representation of preserved data and meta data for the following reasons:

Flexibility. The data model for representing the preserved wiki contents is likely to change over time to meet new requirements and it is not predictable at present how this data model will evolve in the future. In this context, modelling the data with the flexible graph-based RDF data model seems a good choice to us: migrating to a newer data model can be seen as an ontology matching problem for which tools and methods are constantly being developed in Semantic Web research [5, 11].

High semantic expressiveness. In order to read and interpret digital content in the future, it is necessary to preserve its semantics. As a consequence of the continuous evolution of data models, knowledge about data semantics disappears quickly if not specified explicitly [11]. To face this problem, we make use of well-defined standardized Semantic Web vocabularies to define the semantics of our data explicitly.

Existing inference support. Inference enables to find relations between items that were not specified explicitly. By this it is possible to generate additional knowledge about the preserved data that might improve future access to and migration of the data.

Expressive query language. One of the key goals of digital preservation systems is to enable users to re-find and access the data stored in such an archive. This often requires a preservation storage to enable complex and sophisticated queries on its data. Data in RDF graphs can be queried using SPARQL, a rich, expressive, and standardized query language that meets these requirements [8].

Furthermore, we decided to publish the archived data as *Linked Data* [2] in order to exchange them between de-centralized components. Linked Data means that (i) resources are identified using HTTP URIs (ii) de-referencing (i.e., accessing) a URI returns a meaningful representation of the respective resource (usually in RDF) and (iii) these representations include links to other related resources. Data published in this way can easily be accessed and integrated into existing Linked Data.

This highly flexible data representation can be accessed via the Urobe Web interface and can easily be converted to existing preservation meta data standards in order to integrate it with an existing preservation infrastructure.

2.2 A Vocabulary for Describing Wiki Contents.

We have developed an OWL Light vocabulary for the description of wiki core elements by analyzing popular wiki engines as well as common meta data standards from the digital preservation domain and vocabularies from the Semantic Web domain. The core terms of our vocabulary are depicted in Figure 1. Our vocabulary builds upon three common Semantic Web vocabularies: (i) *DCTERMS* for terms maintained by the Dublin Core Metadata Initiative [4], (ii) *SIOC* for describing online communities and their data [5] and (iii) *VoiD* for describing datasets in a Web of Data [6].

The vocabulary was designed to be directly mappable to the PREMIS Data Dictionary 2.0 [9]. We have implemented such a mapping and enable external tools to access the data stored in an Urobe archive as PREMIS XML descriptions [7]. This PREMIS/XML interface makes any Urobe instance a PREMIS-enabled storage that can easily be integrated into other PREMIS-compatible preservation infrastructures.

[3] http://www.w3.org/2004/OWL/

[4] http://purl.org/dc/terms/

[5] http://sioc-project.org/

[6] http://vocab.deri.ie/void/

[7] In compliance to the Linked Data recommendations access to these representations is possible via content negotiation.

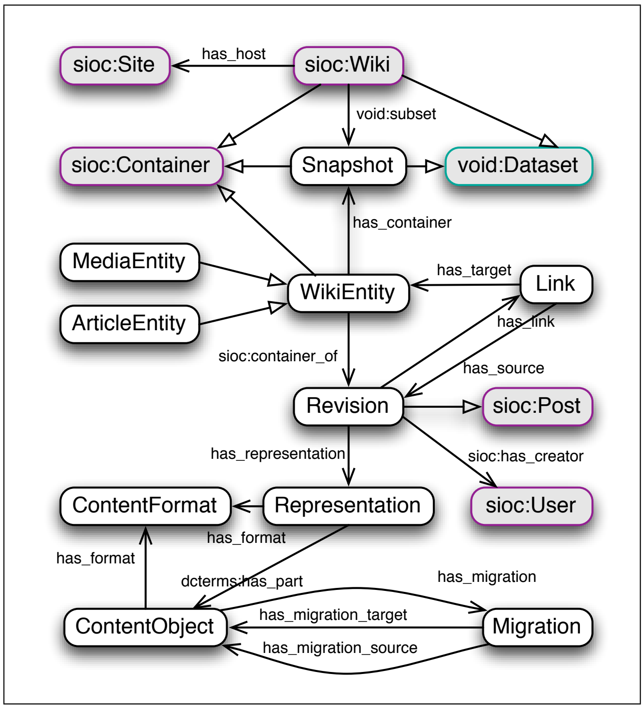


Figure 1: Core terms of the Urobe vocabulary for describing wiki contents. The vocabulary is available at http://urobe-info.mminf.univie.ac.at/vocab.

PREMIS is extensible by design: As RDF graphs can be serialized to XML [8] they are directly embeddable in PREMIS descriptions (using the *objectCharacteristicsExtension* semantic unit). Further, it is possible to describe media objects (i.e., images, videos, documents) that are embedded in wiki pages using appropriate semantic vocabularies like the COMM multimedia ontology [1] (an MPEG-7 based OWL DL ontology that covers most parts of the MPEG-7 standard) or the Music Ontology [9] (that provides a formal framework for describing various music-related information, including editorial, cultural and acoustic information). These descriptions can then be embedded in/mapped to PREMIS descriptions. These meta data could partly be extracted from the object's content itself (e.g., from ID3v2 tags or XMP headers) but also be retrieved directly from the Web of Data (e.g., from the MusicBrainz database, cf. [10]) which could enhance the quality of these meta data considerably.

2.3 Migration of Wiki Articles

Some time ago, the wiki research community started with a first standardization attempt for wiki markup languages [10] which led to a first stable recommendation (Creole 1.0). We therefore decided to implement tools for migrating the source code of wiki articles from their original wiki markup language to Creole 1.0 as soon as they are integrated into our preservation storage. So far we have implemented migration tools for the markup languages of MediaWiki and JspWiki based on components from the WikiModel project [11] .

Creole is a wiki markup that contains common elements of many existing wiki engines. However, it is not able to express all specialized elements that are available in the various markup languages [12] . This means that converting wiki articles to Creole is often a lossy migration step. Therefore Urobe additionally preserves the original article source code in its original markup language to enable less lossy migration in the future. However, some loss is unavoidable in such a migration, although it might concern mostly features of minor importance (such as e.g., specialized coloring of table headings or borders around embedded images). If such features have to be preserved, storing the HTML representation of wiki articles is unavoidable. However, even in this case we consider the preservation of a wiki's core elements as beneficial as it enables integration of the data with other data but also direct reasoning on the archived contents.

Further it is notable, that preserving the article source code instead of its rendered HTML version saves a lot of space in a preservation storage: When we compared the raw byte sizes of HTML and plain source code representations of random Wikipedia articles, we found out that the source code representation uses less than 10% of the HTML size in most cases. Thus, the storage requirements for a Wiki archive could be reduced considerably if the mentioned migration loss is considered acceptable in a particular wiki preservation strategy.

As mentioned before, not only the data themselves but also their semantics that are expressed using our OWL vocabulary will have to be migrated in the future. We have not yet developed tools for the migration of our vocabulary, but are confident that this can be achieved by using tools and methodologies from ontology matching research.

2.4 Modularized, Distributed Architecture

Urobe is a distributed Web application that comprises three central components:

Proxy components access particular wiki implementations, convert their data to RDF and expose these RDF graphs as Linked Data. Proxies know how to access the data stored by a particular wiki software (e.g., by directly interacting with the database the wiki stores its data in).

The format registry stores descriptive and administrative meta data about particular file formats, including descriptions of various wiki markup languages.

The preservation server periodically accesses the proxy components using HTTP requests and harvests all data that were created since the proxy was last accessed. These

[8] `http://www.w3.org/TR/REC-rdf-syntax/`
[9] `http://musicontology.com/`
[10] `http://wiki.wikicreole.org/`
[11] `http://code.google.com/p/wikimodel/`
[12] An example are mathematical formulas in media wiki, see `http://en.wikipedia.org/wiki/Help:Displaying_a_formula`.

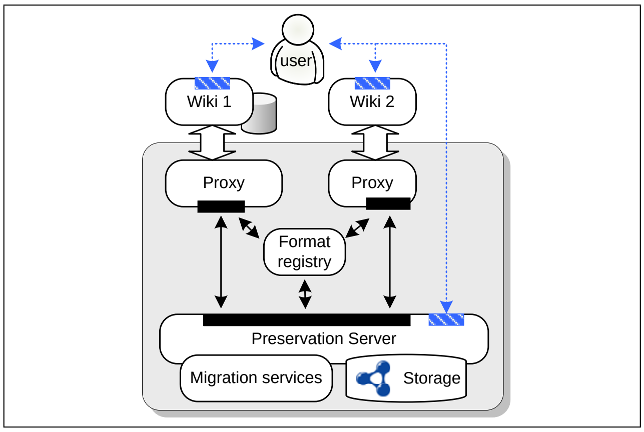


Figure 2: Core architecture of our Urobe prototype. Solid black boxes and arrows denote Linked Data interfaces, dashed blue boxes and arrows denote HTML interfaces.

data are stored in a triple store and interlinked with other data stemming from other wiki instances (e.g., user objects are automatically interlinked when they share a common e-mail address). Such links are then exploited e.g., for data access via the Urobe GUI. A preservation server is able to archive multiple wikis of different types. The internal architecture of the preservation server is influenced by the reference model for an Open Archival Information System (OAIS). Migration services for various object types can be plugged into this server. Currently, object migration is done either immediately after ingestion (for wiki articles) or on demand (for media objects). A preservation workflow component is under development.

The components of Urobe are loosely coupled via HTTP interfaces (cf. Figure 2). This modular architecture and the standardized protocols and formats used by Urobe allow for the easy integration of its components into other applications.

2.5 A Web Interface for Accessing the Urobe Preservation Storage

Human users may access Urobe via an HTML interface (Figure 3) provided by the preservation server component. This interface enables them to search for wiki contents in the Urobe archive using full-text queries and a faceted search approach. Facets for filtering result sets include (i) the wiki(s) the user wants to search, (ii) the time interval the results were created in, (iii) content types and, (iv) the size of multimedia objects. The detail view of articles/media objects presents a timeline of the preserved revisions of this item that indicates all revisions that were created within the search time frame using a different color. Users may navigate to other revisions by simply clicking into the timeline. The original source code of an article as well as all migrated representations are accessible via this screen. A HTML version that is rendered from the preserved Creole source code comprises the default view of an article. Machine actors may further access PREMIS/XML and RDF representations of the stored wiki contents using the Linked Data interface that exposes these data in a machine processable format. The various representation formats are accessible via content negotiation: e.g., when the content type *text/n3* is passed in the Accept header of the HTTP request, Urobe returns a N3-serialized RDF graph describing the respective resource. Urobe also provides a SPARQL endpoint for formulating complex queries over the preservation storage. As future work we further consider to implement a time-based content negotiation mechanism for accessing our preservation storage, as recently presented in [12].

3 CONCLUSIONS

We have formulated requirements and presented a first approach for the digital long-term preservation of wikis, a particular type of a Web 2.0 application. Our approach strongly relies on the adoption of Semantic Web methods and technologies. Wiki contents are modeled using a graph-based data model and their semantics are described using a simple OWL ontology. The advantages of Semantic Web technologies for digital preservation tasks were also recognized by others [7, 8, 3], especially the flexible and extensible way of data representation is considered as beneficial for future data and vocabulary migration as well as for data integration tasks.

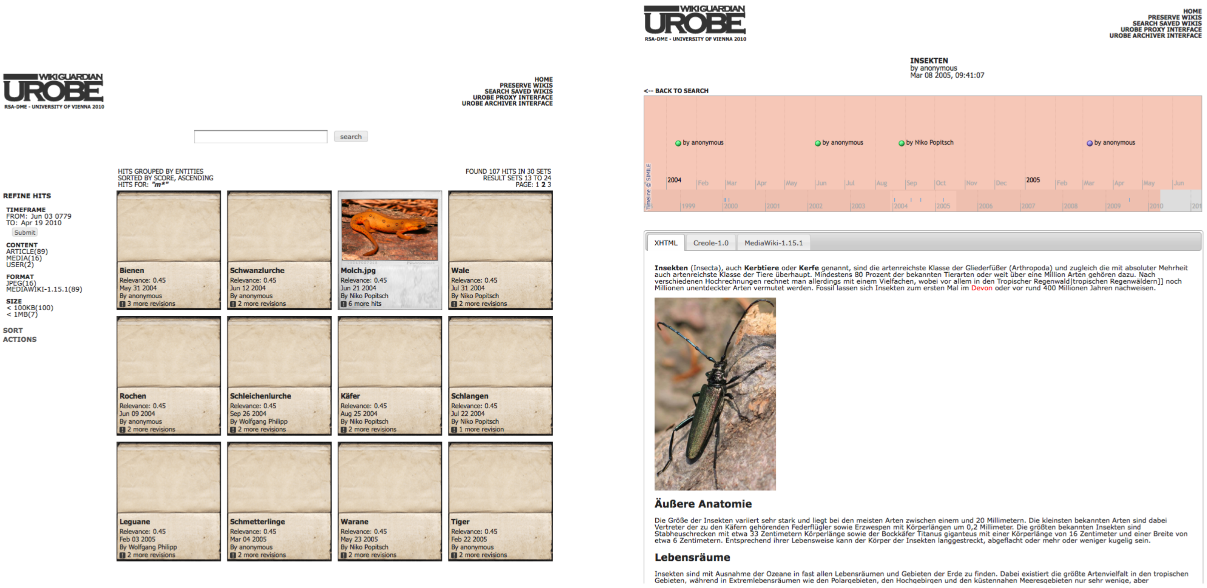
We further contribute a first vocabulary for the abstract description of wiki contents which we consider a precondition for a general wiki preservation strategy. We envision this vocabulary to be continuously improved in the future, which requires algorithms and tools for migrating the data in a Urobe preservation storage to a new vocabulary version. As discussed, we have not yet implemented such a functionality, but due to the strong application of Semantic Web technologies we can benefit directly from the ongoing research in the area of ontology matching.

In the course of the ongoing Urobe project, we aim at extending our vocabulary and implementing support for other semantic vocabularies that are able to capture additional aspects of the preserved data that are of importance in digital preservation, such as context information and provenance meta data.

Finally, our proposed way of exposing the data stored in Urobe as Linked Data enables others to link to these data in a standardized way without compromising their integrity. These externally linked data could then be exploited to harvest additional preservation meta data and ultimately to improve future content migration steps. Further, others could directly benefit from the invariant data in such an archive by being able to create stable links to particular revisions of wiki core elements.

4 ACKNOWLEDGEMENT

This research is partly funded by the *Research Studios Austria* program of the Federal Ministry of Economy, Family and Youth of the Republic of Austria (BMWFJ).



HITS GROUPED BY ENTITIES
SORTED BY SCORE, ASCENDING
HITS FOR: *"m*"*

FOUND 107 HITS IN 30 SETS
RESULT SETS 13 TO 24
PAGE: 1 **2** 3

TIMEFRAME
FROM: Jun 03 0779
TO: Apr 19 2010

CONTENT
ARTICLE(89)
MEDIA(16)
USER(2)

FORMAT
JPEG(16)
MEDIAWIKI-1.15.1(89)

SIZE
< 100KB(100)
< 1MB(7)

Bienen
Relevance: 0.45
May 31 2004
By anonymous
3 more revisions

Schwanzlurche
Relevance: 0.45
Jun 12 2004
By anonymous
2 more revisions

Molch.jpg
Relevance: 0.45
Jun 21 2004
By Niko Popitsch
6 more hits

Wale
Relevance: 0.45
Jul 31 2004
By Niko Popitsch
2 more revisions

Rochen
Relevance: 0.45
Jun 09 2004
By anonymous
2 more revisions

Schleichenlurche
Relevance: 0.45
Sep 26 2004
By Wolfgang Philipp
2 more revisions

Käfer
Relevance: 0.45
Aug 25 2004
By Niko Popitsch
2 more revisions

Schlangen
Relevance: 0.45
Jul 22 2004
By Niko Popitsch
1 more revision

Leguane
Relevance: 0.45
Feb 03 2005
By Wolfgang Philipp
2 more revisions

Schmetterlinge
Relevance: 0.45
Mar 04 2005
By anonymous
2 more revisions

Warane
Relevance: 0.45
May 23 2005
By Niko Popitsch
2 more revisions

Tiger
Relevance: 0.45
Feb 22 2005
By anonymous
2 more revisions

(a) Main search screen.

INSEKTEN
by anonymous
Mar 08 2005, 09:41:07

Insekten (Insecta), auch **Kerbtiere** oder **Kerfe** genannt, sind die artenreichste Klasse der Gliederfüßer (Arthropoda) und zugleich die mit absoluter Mehrheit auch artenreichste Klasse der Tiere überhaupt. Mindestens 80 Prozent der bekannten Tierarten oder weit über eine Million Arten gehören dazu. Nach verschiedenen Hochrechnungen rechnet man allerdings mit einem Vielfachen, wobei vor allem in den Tropischer Regenwald|tropischen Regenwäldern]] noch Millionen unentdeckter Arten vermutet werden. Fossil lassen sich Insekten zum ersten Mal im Devon oder vor rund 400 Millionen Jahren nachweisen.

Äußere Anatomie

Die Größe der Insekten variiert sehr stark und liegt bei den meisten Arten zwischen einem und 20 Millimetern. Die kleinsten bekannten Arten sind dabei Vertreter der zu den Käfern gehörenden Federflügler sowie Erzwespen mit Körperlängen um 0,2 Millimeter. Die größten bekannten Insekten sind Stabheuschrecken mit etwa 33 Zentimetern Körperlänge sowie der Bockkäfer Titanus giganteus mit einer Körperlänge von 16 Zentimeter und einer Breite von etwa 6 Zentimetern. Entsprechend ihrer Lebensweise kann der Körper der Insekten langgestreckt, abgeflacht oder mehr oder weniger kugelig sein.

Lebensräume

Insekten sind mit Ausnahme der Ozeane in fast allen Lebensräumen und Gebieten der Erde zu finden. Dabei existiert die größte Artenvielfalt in den tropischen Gebieten, während in Extremlebensräumen wie den Polargebieten, den Hochgebirgen und den küstennahen Meeresgebieten nur sehr wenige, aber

(b) Detail view.

Figure 3: Urobe graphical user interface. The left screenshot shows the main search screen, including the full-text search and the faceted search interface. The right screenshot shows the detail view of a preserved article: the timeline on top of the screen visualizes the revisions of the corresponding wiki article. Below, various representations of the article (XHTML, Creole, original markup) can be accessed.

5 REFERENCES

[1] Richard Arndt, Raphael Troncy, Steffen Staab, Lynda Hardman, and Miroslav Vacura. COMM: Designing a well-founded multimedia ontology for the web. In *Proceedings of the ISWC '07*, pp. 30–43, 2007.

[2] Tim Berners-Lee Christian Bizer, Tom Heath. Linked data - the story so far. *IJSWIS*, 5(3):1–22, 2009.

[3] Laura E. Campbell. Recollection: Integrating data through access. In *Proceedings of the ECDL '09*, pp. 396–397, 2009.

[4] Jeremy J. Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. Named graphs, provenance and trust. In *Proceedings of the WWW '05*, pp. 613–622, 2005.

[5] Giorgos Flouris, Dimitris Manakanatas, Haridimos Kondylakis, Dimitris Plexousakis, and Grigoris Antoniou. Ontology change: Classification and survey. *Knowl. Eng. Rev.*, 23(2):117–152, 2008.

[6] Mark Greaves. Semantic web 2.0. *IEEE Intelligent Systems*, 22(2):94–96, 2007.

[7] Jane Hunter and Sharmin Choudhury. Panic: an integrated approach to the preservation of composite digital objects using semantic web services. *Int. J. Digit. Libr.*, 6(2):174–183, 2006.

[8] Gautier Poupeau and Emmanuelle Bermès. Semantic web technologies for digital preservation : the spar project. In *Proceedings of the Poster and Demonstration Session at ISWC2008*, 2008.

[9] PREMIS Editorial Committee. Premis data dictionary for preservation metadata, version 2.0, 2008. `http://www.loc.gov/standards/premis/`.

[10] Yves Raimond, Christopher Sutton, and Mark Sandler. Interlinking music-related data on the web. *IEEE MultiMedia*, 16(2):52–63, 2009.

[11] Christoph Schlieder. Digital Heritage: Semantic Challenges of Long-term Preservation. *submitted to the Semantic Web Journal (SWJ)*, 2010. `http://www.semantic-web-journal.net/content/new-submission-digital-heritage-semantic-challenges-long-term-preservation`.

[12] Herbert Van de Sompel, Robert Sanderson, Michael Nelson, Lyudmila Balakireva, Harihar Shankar, and Scott Ainsworth. An HTTP-based Versioning Mechanism for Linked Data. *LDOW2010, Co-located with WWW '10*, 2010.

[13] W3C Semantic Web Activity - RDF Data Access Working Group. Sparql query language for rdf. Technical report, W3C, 2008.