Publishing Provenance Information on the Web using the Memento Datetime Content Negotiation

Sam Coppens
Ghent University - IBBT
Multimedia Lab
Gaston Crommenlaan 8 /201
Ghent, Belgium
sam.coppens@ugent.be

Erik Mannens
Ghent University - IBBT
Multimedia Lab
Gaston Crommenlaan 8 /201
Ghent, Belgium
erik.mannens@ugent.be

Davy Van Deursen
Ghent University - IBBT
Multimedia Lab
Gaston Crommenlaan 8 /201
Ghent, Belgium
davy.vandeursen@ugent.be

Patrick Hochstenbach
Boekentoren - Ghent
University Library
Rozier 9
Ghent, Belgium
patrick.hochstenbach@ugent.be

Bart Janssens
Descartes Systems Group
Duwijckstraat 17
lier, Belgium
bjanssens@descartes.com

Rik Van de Walle
Ghent University - IBBT
Multimedia Lab
Gaston Crommenlaan 8 /201
Ghent, Belgium
rik.vandewalle@ugent.be

ABSTRACT

In Belgium, we developed a digital long-term preservation archive to preserve the information from our heritage institutions. This platform harvests the information from the institutions, preserves the information for the long term and disseminates the information as Linked Open Data. Our platform produces many different versions of the harvested data to keep the information accessible over time when, e.g., mapping the metadata or transcoding the multimedia files, but it also produces a lot of provenance information relating all those different versions of a resource. For publishing this information as Linked Open Data, we extended our Linked Open Data server with Memento datetime content negotiation. Next to this, we extended the Memento framework to also publish the provenance information of those datetime content negotiated versions using an HTTP provenance link header for automatic discovery of the provenance information. This way, our framework allows to publish the information of a resource as Linked Open Data, including all its previous versions and their provenance information, in a web-accessible manner.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: General

General Terms

Design, Management, Standardization

Keywords

Linked Open Data, Memento datetime content negotiation, Provenance

1. INTRODUCTION

Many organisations and private persons still possess a lot of material which is stored on analogue carriers. This material is mostly part of important cultural heritage anywhere. At this moment, the analogue carriers are degrading and continuously losing quality, making the data inaccessible. While we are still able to see wall paintings from millennia ago, many documents from merely a decade or two decades ago have become inaccessible, e.g., *WordPerfect* files. Some refer to this situation as the *Digital Dark Age*[4]. Digital long-term preservation forms the solution for this issue. A digital long-term archive has the necessary processes in place to withstand many long-term preservation risks, e.g., bitrot, file formats becoming obsolete, etc. These preservation processes make sure the content remains intact and accessible over time.

The project *Archipel*[1] initiates the dissemination and digital long-term preservation of the cultural heritage in Flanders, Belgium, and researches the problems encountered with digital long-term preservation. In this project, we developed a platform that harvests data coming from various institutions (libraries, archival institutions, the art sector (museums), and the broadcasters), preserves the data for the long term and disseminates the data as Linked Open Data [1] (LOD) Dublin Core[2] records.

To guarantee the long-term preservation of the harvested content, our platform has the necessary processes in place to keep the information intact and interpretable, in line with the Open Archival Information System (OAIS) reference model [5] for the long-term preservation of information. These processes rely heavily on the provenance information of the harvested data, but at the same time produce also a lot of provenance information. This provenance information is modelled using a semantic implementation of the PREMIS 2.0 data dictionary[3], i.e., PREMIS OWL[4].

Our developed platform generates many different versions of the harvested data, i.e., metadata and referenced multimedia files, via its preservation processes. These resources, their previous versions and their provenance information, re-

LDOW2011, March 29, 2011, Hyderabad, India.

[1] http://www.archipelproject.be
[2] http://dublincore.org/
[3] http://www.loc.gov/standards/premis/
[4] http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/index.html

lating the different versions, will be published on the Web as LOD. When preserving information for the long term and publishing the information as LOD at the same time, different problems arise. First of all, we need to have persistent URIs for our resources, which will publish the information of a certain version of the resource. Another problem involves the enrichments that occur on the resources before publishing them as LOD. These enrichments will not always remain valid over time. We need a way for preserving the temporality of these enrichments. The last problem being tackled in this paper is the publication of the provenance information on the Web which will allow automatic discovery of the provenance information.

To solve these problems, our developed platform is extended with the Memento[5] [13] datetime content negotiation. This datetime content negotiation will allow to select the appropriate version, called memento in the Memento framework, of the archived information and to publish it on a persistent URI. This datetime content negotiation will also solve the problem of preserving the temporality of the enrichments of the archived information. The different versions of the archived information are linked to each other via their provenance information. To publish the provenance information of each version on the Web, we extended the Memento framework to offer provenance links using a special Hypertext Transfer Protocol (HTTP)[8] link header for automatic discovery of the provenance information.

In this paper, we present how our digital long-term preservation platform is able to publish the provenance information on the Web. First, Section 2 describes some related work on this topic. Then, in Section 3, we introduce our semantic layered metadata model, which allows the archive to deal with the diversity of metadata records coming from diverse institutions and to track the provenance of the harvested data. Section 4 describes the distributed architecture of the archive and its processes. Section 5 explains the publication of the content and its provenance information using the Memento framework, extended to provide provenance information. We end with a conclusion in Section 6.

2. RELATED WORK

Interest in digital preservation can be seen by the multitude of projects in this area. Planets (Preservation and Long-term Access through Networked Services)[6] was especially aimed at defining guidelines for preservation planning. However, it did not tackle the integration of different existing metadata formats, or the dissemination of the metadata as LOD. Likewise, the Prestospace (Preservation towards storage and access) project's objective was to provide technical solutions and integrated systems for a complete digital preservation of all kinds of audio-visual collections [7]. The project was especially focussed on the underlying technologies, e.g., automated generation of metadata or detection of errors in content [11], but without using a standardised, semantic preservation model to support the archiving, nor do they tackle the problem of publishing the generated provenance information to the Web.

The CASPAR project (Cultural Artistic and Scientific knowledge for Preservation, Access, and Retrieval) presented technologies for digital preservation [8]. The OAIS Reference Model was chosen as the base platform, and the project was focused on implementing the different steps in the preservation workflow. They focus more on preservation services than on describing the preservation information. BOM Vlaanderen [9], a national research project, was aimed at preservation and disclosure of audio-visual content in Flanders. Additionally, it looked at ways to unify different metadata standards currently used for describing audio-visual content. Current trends are on integrating different media archives. PrestoPRIME researches and develops practical solutions for the long-term preservation of digital media objects, programmes and collections, and finds ways to increase access by integrating the media archives with European on-line digital libraries in a digital preservation framework [10].

The previous discussed related work were focusing on the digital long-term preservation, not on the more general problem of enabling their provenance information on the Web. For the work done is this area, the work of the W3C Provenance Incubator Group[11] is the major reference. This incubator group produced working definitions for provenance information, provided a state-of-the-art understanding and developed a roadmap for development and possible standardisation of provenance on the Web. This work included defining key dimensions for provenance, collecting use cases, designing three flagship scenarios from the use cases, creating mappings between existing provenance vocabularies, looking how provenance could fit in the Web architecture and providing a state-of-the-art report on the current provenance activities. Their work is summarised in a final report [6]. The first flagship scenario describes a news aggregator site that assembles news items from a variety of data sources, e.g., news sites, blogs and tweets. The provenance records of these data providers can help with verification, credit and licensing. This flagship scenario could be covered by publishing the provenance information using our framework. What still forms a problem is the lack of a standardised metadata model for publishing provenance on the Web. In our framework, we publish the provenance information as Linked Open Data using PREMIS OWL. This information is only interoperable in the long-term preservation context, where PREMIS is well known, not in a Web context. This standardised provenance model for the Web is still a major research area. The work of the W3C Provenance Incubator Group was a first step into that direction.

Another interesting work done in the area of publishing provenance for linked data is the paper of Olaf Hartig and Jun Zhao published at IPAW [7]. In that paper they describe the Provenance Vocabulary[12] used for describing the provenance information as Linked Open Data. Next to this, they also offer ways of publishing this provenance information for Linked Data. They discuss how provenance can be added to Linked Data objects, how provenance can be included into RDF dumps and how the provenance information can be queried using SPARQL endpoints. This work enables provenance for Linked Data, but it does not offer solutions for automatic discovery of the provenance infor-

[5] http://www.mementoweb.org
[6] http://www.planets-project.eu/
[7] http://prestospace.org/project/index.nl.html
[8] http://www.casparpreserves.eu/
[9] https://projects.ibbt.be/bom-vl
[10] http://www.prestoprime.org/
[11] http://www.w3.org/2005/Incubator/prov/wiki/W3C_Provenance_Incubator_Group_Wiki
[12] http://purl.org/net/provenance/

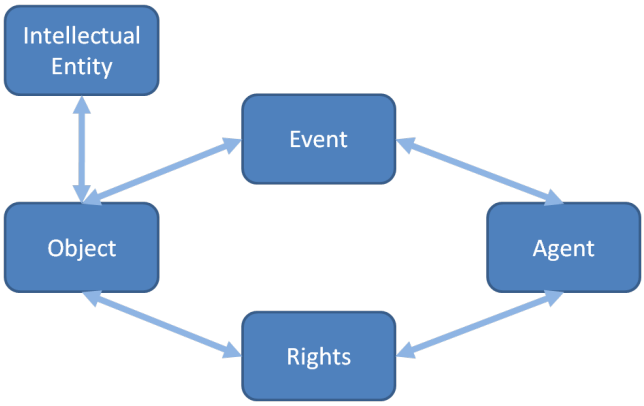

Figure 1: Data Model of the Premis 2.0 Data Dictionary.

mation or ways for publishing provenance on the Web beyond using semantic web technologies. Future work could involve publishing the provenance information using this vocabulary, which is more suited for publication on the Web than PREMIS OWL, which is intended to be a data model for digital long-term archives. The mapping table, relating various provenance vocabularies, produced by the W3C Incubator Group[13] will be the reference for this work.

3. LAYERED METADATA MODEL

Descriptive metadata schemes describe the content of the harvested data: subject, author, date of creation, file format, etc. This metadata makes it possible to manage and search the complete digital archive. When archiving data coming from different sectors like the broadcast sector, the libraries, the cultural sector, and the archival sector, a problem arises concerning descriptive metadata. Many of the institutions already have descriptive metadata using domain-specific metadata models. To deal with this diversity of metadata models, the descriptive metadata is mapped to Dublin Core RDF [12] and is archived along with the data in their original metadata format, e.g., MARC , so there is no information loss. This gives the archive the necessary tools to search the whole archive. When finding the data of interest, the original metadata that is stored as data can still be presented to the users.

DC RDF was chosen as format for the descriptive metadata, as it is a broadly accepted descriptive schema. The power of this schema is its simplicity and generality. It only consists of fifteen fields among which creator, subject, coverage, description, and date. It can answer to the basic questions: Who, What, Where, and When. All the fields in DC are optional and repeatable. This makes it possible to map relatively easily almost all the descriptive metadata schemes to DC RDF as many institutions already support DC. This choice will also benefit the publication of the diverse records coming from the institutions as LOD, as will be discussed in Section 5.

To store the preservation metadata, we developed a semantic binding of the PREMIS 2.0 Data Dictionary. The

[13]`http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings`

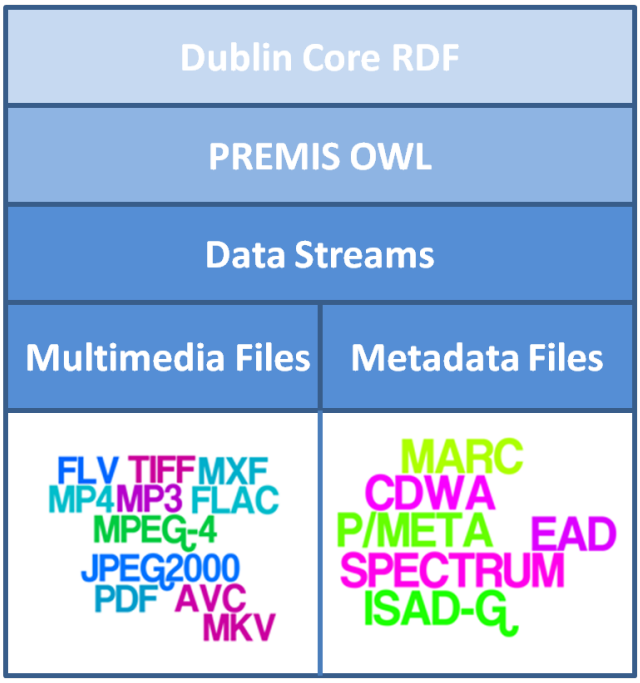

Figure 2: Layered data model for the long-term archive.

PREMIS 2.0 Data Dictionary was especially designed for storing provenance information in the context of digital long-term preservation and is in line with the requirements of OAIS. This PREMIS OWL schema is currently undergoing a standardisation process and will soon be published on a more stable URL of the Library of Congress. The PREMIS 2.0 Data Dictionary is described by a data model, which consists of five semantic units or classes important for digital preservation purposes:

- *Intellectual Entities*: a part of the content that can be considered as an intellectual unit for the management and the description of the content. This can be for example a book, a photo, or a database.
- *Object*: a discrete unit of information in digital form, typically multimedia objects related to the intellectual entity.
- *Event*: An action that has an impact on an object or an agent.
- *Agent*: a person, institution, or software application that is related to an event of an object or is associated to the rights of an object.
- *Rights*: description of one or more rights, permissions of an object or an agent.

Intellectual entities, events, and rights are directly related to an object, whereas an agent can only be related to an object through an event or through rights, as can be seen on Figure 1. This way, not only the changes to an object are stored, but the event involved in this change is also described. These relationships offer the necessary tools to properly store the provenance of an archived object. The

rights metadata needed for preservation are covered by the rights entity, which relates to the agent entity and the object entity. The binary metadata, technical metadata and structural metadata are encapsulated in the PREMIS data dictionary via the description of the object entity. Examples of an PREMIS OWL Object entity, Event entity, Rights entity and Agent entity are given in the resp. Listing 1, Listing 2, Listing 3, and Listing 4.

```
@prefix rdf:          <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:         <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:          <http://www.w3.org/2002/07/owl#> .
@prefix premisowl:    <http://multimedialab.elis.ugent.be/users/samcoppe/
      ontologies/Premis/premis.owl#>.

<object1>                          a              premisowl:File;
       premisowl:preservationLevel                 <object1PreservationLevel>;
       premisowl:significantProperties             <object1SignificantProperties>;
       premisowl:objectCharacteristics             <object1ObjectCharacteristics>;
       premisowl:originalName                      "0001h.tif";
       premisowl:storage                           <object1Storage>;
       premisowl:environment                       <object1Environment>;
       premisowl:linkingEvent                      <event2>;
       premisowl:linkingRightsStatement            <rightsstatement1>;
       premisowl:linkingIntellectualEntity         <dublinCoreDescription1>.

<object1PreservationLevel>      a                  premisowl:PreservationLevel;
       premisowl:preservationLevelValue            "0";
       premisowl:preservationLevelRole             "master_copy";
       premisowl:preservationLevelDateAssigned     "2010-07-29T14:41:28".

<object1SignificantProperties>  a                  premisowl:SignificantProperties
      ;
       premisowl:significantPropertiesType         "behavior";
       premisowl:significantPropertiesValue        "hyperlinks_traversable".

<object1ObjectCharacteristics>  a                  premisowl:ObjectCharacteristics
      ;
       premisowl:compositionLevel                  "0";
       premisowl:fixity                            <object1Fixity>;
       premisowl:size                              "20800896";
       premisowl:format                            <object1Format>;
       premisowl:creatingApplication               <object1CreatingApplication1>;
       premisowl:objectCharacteristicsExtension<
              object1CharacteristicsExtension>.

<object1Fixity>                    a              premisowl:Fixity;
       premisowl:messageDigestAlgorithm            "MD5";
       premisowl:messageDigest
              b03197ad066cd719906c55eb68ab8d";
       premisowl:messageDigestOriginator           "LocalDCMS".

<object1Format>                    a              premisowl:Format;
       premisowl:formatDesignation                 <object1FormatDesignation>;
       premisowl:formatRegistry                    <object1FormatRegistry>.

<object1FormatDesignation>         a              premisowl:FormatDesignation;
       premisowl:formatName                        "image/tiff";
       premisowl:formatVersion                     "6.0".

<object1FormatRegistry>            a              premisowl:FormatRegistry;
       premisowl:formatRegistryName                "PRONOM";
       premisowl:formatRegistryKey                 "fmt/10";
       premisowl:formatRegistryRole                "specification".

<object1CreatingApplication1>      a              premisowl:CreatingApplication;
       premisowl:creatingApplicationName           "Adobe_Photoshop";
       premisowl:creatingApplicationVersion        "CS2";
       premisowl:dateCreatedByApplication          "2006-09-20T08:29:02".

<object1Storage>                   a              premisowl:Storage;
       premisowl:contentLocation                   <object1ContentLocation>;
       premisowl:storageMedium                     "disk".

<object1ContentLocation>           a              premisowl:ContentLocation;
       premisowl:contentLocationType               "filepath";
       premisowl:contentLocationValue              "amserver".

<object1Environment>               a              premisowl:Environment;
       premisowl:environmentCharacteristic         "recommended";
       premisowl:environmentPurpose                "render";
       premisowl:environmentPurpose                "edit";
       premisowl:software                          <object1Software1>;
       premisowl:hardware                          <object1Hardware1>.

<object1Software1>                 a              premisowl:Software;
       premisowl:swName                            "Adobe_Acrobat";
       premisowl:swVersion                         "5.0";
       premisowl:swType                            "renderer".

<object1Hardware1>                 a              premisowl:Hardware;
       premisowl:hwName                            "Intel_x86";
       premisowl:hwType                            "processor";
       premisowl:hwOtherInformation                "60_mhz_minimum".
```

Listing 1: PREMIS OWL Object Instance in N3 Notation.

```
@prefix rdf:          <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:         <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:          <http://www.w3.org/2002/07/owl#> .
@prefix premisowl:    <http://multimedialab.elis.ugent.be/users/samcoppe/
      ontologies/Premis/premis.owl#>.

<event1>                                       a       premisowl:Event;
       premisowl:eventIdentifier                       <event1ID>;
       premisowl:eventType                             "dissemination_migration";
       premisowl:eventDateTime                         "2010-08-06T00:00:00.002";
       premisowl:eventDetail                           "ImageMagick";
       premisowl:eventOutcomeInformation               <event1OutcomeInformation>;
       premisowl:linkingAgent                          <agent1>;
       premisowl:linkingObject                         <object1>;
       premisowl:linkingObject                         <object2>;
              .

<event1ID>                                     a       premisowl:EventIdentifier;
       premisowl:identifierType                        "LocalDCMS";
       premisowl:identifierValue                       "E002.1";
       .

<event1OutcomeInformation>                     a
      premisowl:EventOutcomeInformation;
       premisowl:eventOutcome                          "successful";
       .
```

Listing 2: PREMIS OWL Event instance in N3 notation.

Employing a data model with the original metadata, the mapped Dublin Core RDF descriptions and the PREMIS OWL metadata for storing the provenance leads to a layered, semantic metadata model, which the archive uses for management, dissemination and preservation purposes, as depicted in Figure 2.

4. ARCHITECTURE

In this section, our architecture of the digital long-term preservation archive is described. In this networked world, various resources are linked to each other. We do not want to build yet another central e-depot, but a distributed network of storage components. For this reason, the platform will have a service oriented architecture[14] (SOA). This SOA will make use of a central service hub, which will offer the needed services for the platform. The objectives of our platform are twofold:

- Disseminate the content and provenance information as LOD.
- Enable long-term preservation.

Our architecture is depicted in Figure 3. The green arrow indicates the dissemination path, the red arrow stipulates the preservation path. The basic components of our architecture are:

- *Repositories*: these are the repositories of the diverse institutions, which have their content published online, using the OAI-PMH protocol [10], depicted in Figure 3 in box 1.
- *Shared Repositories*: for those institutions, which do not have published their content on-line, our Archipel project foresees several shared repositories, using *Omeka*[15] or *MediaMosa*[16], which will publish their content online using the OAI-PMH protocol. This is shown in Figure 3 in box 2.

[14] http://opengroup.org/projects/soa/
[15] http://omeka.org/
[16] http://www.mediamosa.org/

```
@prefix rdf:            <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:           <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:            <http://www.w3.org/2002/07/owl#> .
@prefix premisowl:      <http://multimedialab.elis.ugent.be/users/samcoppe/
      ontologies/Premis/premis.owl#>.

<rights1>                                   a       premisowl:License;
       premisowl:rightsStatementIdentifier          <rights1ID>;
       premisowl:rightsBasis                        "license";
       premisowl:licenseInformation                 <licenseInformation1>;
       premisowl:rightsGranted                      <rightsGranted1>;
       premisowl:linkingObject                      <object1>;
       premisowl:linkingObject                      <object2>;
       premisowl:linkingAgent                       <>;
            .

<rights1ID>                                 a
      premisowl:RightsStatementIdentifier;
       premisowl:identifierType                     "URL";
       premisowl:identifierValue                    "http://archipellod.demo.ibbt.
            be:8080/rights/resource/dissemination";
       .

<licenseInformation1>                       a       premisowl:LicenseInformation;
       premisowl:licenseIdentifier                  <license1identifier>;
       premisowl:licenseTerms                       "Here_comes_the_actual_text_of_
            the_license._(under_development)";
       premisowl:licenseNote                        "These_objects_may_be_
            disseminated.";
       .

<license1identifier>                        a       premisowl:LicenseIdentifier;
       premisowl:identifierType                     "URL";
       premisowl:identifierValue                    "http://archipellod.demo.ibbt.
            be:8080/license/resource/dissemination";
       .

<rightsGranted1>                            a       premisowl:LicenseInformation;
       premisowl:act                                <license1identifier>;
       premisowl:termOfGrant                        <license1termofgrant>;
       .

<license1termofgrant>                       a       premisowl:TermOfGrant;
       premisowl:startDate                          "2009-09-01T08:30:00";
       .
```

Listing 3: PREMIS OWL Rights instance in N3 notation.

- *Integration Server*: this server provides an integration layer for orchestrating all the needed processes, which are all implemented as web services, e.g., transcoding services. Box 3 of Figure 3 shows this.
- *LOD server*: this server is used for the dissemination of the content and the provenance information, with a triple store as a storage back-end, shown in box 4 of Figure 3.
- *CMS*: The CMS will store the archived content, using persistent identifiers and cloud storage, depicted in box 5 of Figure 3. For this *Fedora Commons*[17] is used.

[17] `http://fedora-commons.org/`

```
@prefix rdf:            <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:           <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:            <http://www.w3.org/2002/07/owl#> .
@prefix premisowl:      <http://multimedialab.elis.ugent.be/users/samcoppe/
      ontologies/Premis/premis.owl#>.

<agent1>                                    a       premisowl:Event;
       premisowl:agentIdentifier                    <agent1ID>;
       premisowl:agentType                          "person";
       premisowl:agentName                          "Sam_Coppens";
       premisowl:linkingAgent                       <agent1>;
       premisowl:linkingObject                      <object1>;
       premisowl:linkingObject                      <object2>;
            .

<agent1ID>                                  a       premisowl:AgentIdentifier;
       premisowl:identifierType                     "OpenID";
       premisowl:identifierValue                    "http://smcoppens.
            archipelopenID.be";
       .
```

Listing 4: PREMIS OWL Agent instance in N3 notation.

- *Identity Service*: with this distributed architecture an identity server is needed for authentication across the different systems, shown in box 6 of Figure 3.

For building our distributed, digital long-term preservation platform, we need an integration server to orchestrate the different processes, based on SOA technology. An Enterprise Service Bus (ESB) provides the open, standards-based connectivity infrastructure for the service oriented architecture and allows these services to exchange data with one another as they participate in our processes. Orchestration between services is handled by a workflow engine. This engine is integrated in the service bus architecture and supports the execution of the preservation processes. An executable preservation process is defined by a control flow that consists of a combination of basic and structured activities. For the communication, the 'Simple Object Access Protocol' (SOAP)[2] is used, a protocol specification for exchanging structured information between services. This integration server is built using the *Porthus*[18] *.NET* Integration server.

The whole preservation/dissemination cycle starts with a **harvesting process**, which will harvest the metadata, and the referenced files. The metadata harvested, is described using several descriptive metadata formats, e.g., MARC, DC, or CDWA. For management and dissemination purposes this metadata needs to be mapped to DC RDF. For this, we rely on a **mapping service**, which will map the incoming metadata to DC descriptions.

If the content also to be preserved, the original metadata record, the mapped DC RDF record and the referenced files get packed into a Submission Information Package (SIP), according to the OAIS specifications by the **SIP creator service**. For this SIP, the *BagIt* [3] package format is used. This SIP package is then delivered to the CMS, using the **SIP ingest service**.

When ingesting this *BagIt* package into the CMS, it has to be supplemented with the preservation information to form an Archival Information Package (AIP) in the OAIS terminology. This package holds all the different versions of the metadata and the multimedia files, referenced by the metadata files. For this preservation information, we will use our PREMIS OWL ontology. During this ingest process, all files in the package get a PREMIS *Object* description, related to the mapped DC RDF description, thus becoming the PREMIS *intellectual entity*. For this we rely on a **characterisation service**, which will identify the file format of the files and model the files as PREMIS *Objects*. Every action performed on such a PREMIS *Object*, will get related to that *Object* and will be modeled as a PREMIS *Event*. This way, the platform is able to store and track the provenance of the descriptive metadata and the referenced multimedia files.

The next thing within the workflow is the migration of the stored, related multimedia files. These files get migrated to a file format, defined by the archives preservation plans. Such a preservation plan can stipulate, e.g., that all image files must be migrated to the TIFF file format to keep the image information accessible for long-term preservation purposes, or, e.g., that all image files must be migrated to the JPEG file format to keep the image information accessible for dissemination purposes. For this, we need **migration services**, which can then migrate various incoming file for-

[18] `http://www.porthus.be`

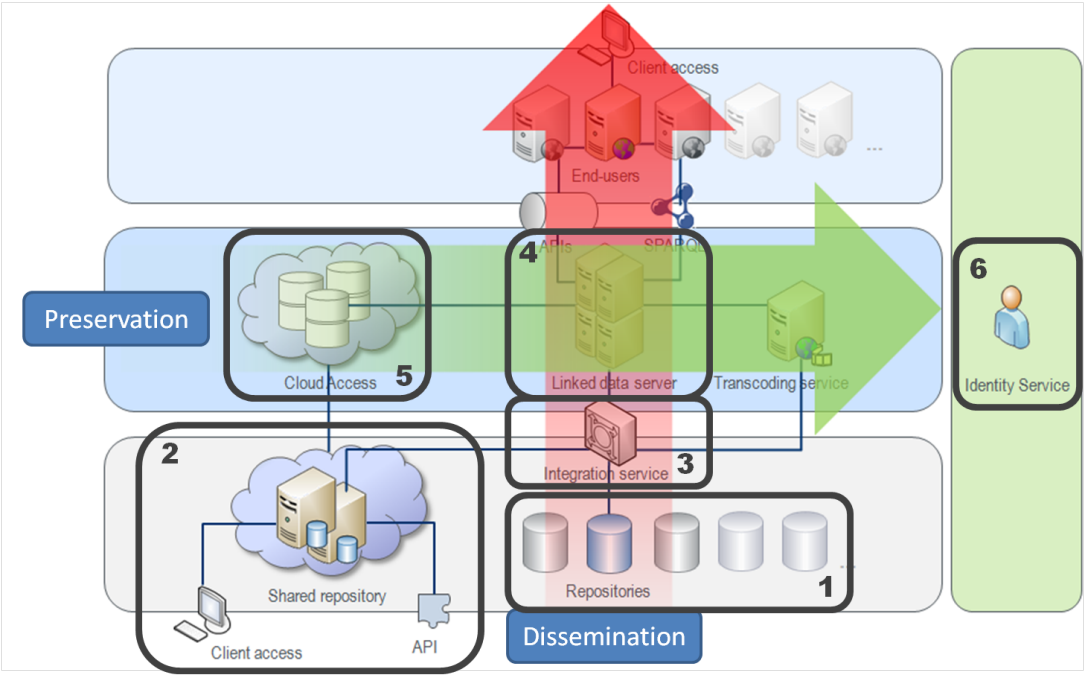

Figure 3: Architecture of the long-term preservation platform.

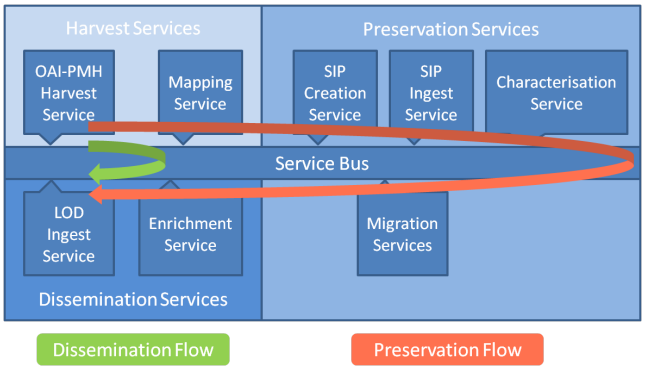

Figure 4: Schematic Overview of the Service Bus and its Connected Services.

mats to the appropriate file format according the preservation plans. This migration will extend the AIP package with the extra migrated data stream. This data stream is then passed to the characterisation service to get a PREMIS *Object* description of the generated data stream and the preservation information is also extended with a description of the migration service as a PREMIS *Event* relating the source object to the migrated object.

During the last phase, the archived information is moved to the LOD server for dissemination of the information. For this, the descriptive DC RDF metadata will get enriched by the **enrichment service** before it gets ingested into the LOD server's triple store by the **LOD ingest service**. For the enrichment service, the platform relies on data sources like the *OpenCalais* infrastructure[19] for extracting these named entities, *GeoNames*[20] for enriching the locations, *DBPedia*[21] for enriching the persons, organisations and events, *BibNet*[22] for authors, singers and music bands enrichment, and *Toerisme Vlaanderen*[23] for touristic information enrichment on locations. This way, our approach provides *i)* unique identifiers for the resource and *ii)* formalised knowledge about this resource.We will not only disseminate the intellectual entity, i.e., the descriptive metadata, but also the preservation information, so the end-user has access to all the information available about that object.

If the harvested content does not need to be preserved, it is directly routed to our **enrichment service**, which will interlink the data with external data sources after harvesting and mapping the metadata. This enriched DC description then gets ingested into the triple store of the LOD server, which automatically publishes the enriched DC records as LOD.

5. PUBLICATION

Our architecture, described in the previous section, ingests all the harvested and generated information into our triple store. This information, including the provenance information, needs to be disseminated as Linked Open Data. For this dissemination, we want to have stable URIs [9], e.g., `http://../record/VTi/1/oai:archipel1.demo.ibbt.be:10` for the harvested original resources. These resources change over time via the preservation processes. Every version of

[19] `http://www.opencalais.com/`
[20] `http://www.geonames.org`
[21] `http://dbpedia.org`
[22] `http://www.bibnet.be/`
[23] `http://www.toerismevlaanderen.be`

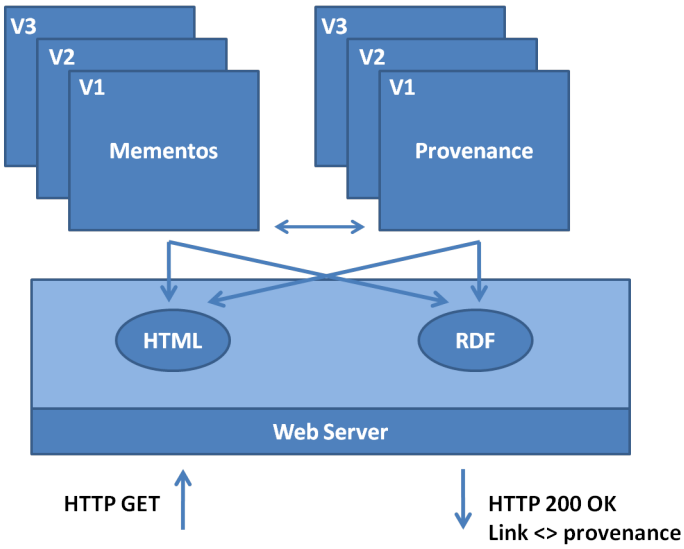

Figure 5: Schematic Overview of the Content Negotiation.

the resource has another URI, e.g., `http://../record/VTi/1/oai:archipel1.demo.ibbt.be:10_V3`. To link from the original resource with a stable URI to the appropriate version URI, we extended our Linked Open Data server with the Memento datetime content negotiation[24], besides the mediatype content negotiation. This mechanism allows the platform to publish the information on persistent URIs. Based on the Memento datetime content negotiation the right version of that resource is selected and published as LOD. This mechanism is depicted in Figure 5 and explained in publication [14].

5.1 Memento Datetime Content Negotiation

The Memento framework is based on HTTP and HTTPS URIs and introduces several concepts:

- Original Resource (URI-R): This resource is archived for the long-term and has several versions.
- Memento (URI-Mj): This resource refers to one of the versions of an original resource.
- TimeGate (URI-G): The TimeGate for an original resource is a resource that supports the datetime content negotiation.
- TimeMap (URI-T): A TimeMap for an original resource lists the URIs of all the mementos of that original resource.

The Memento framework is based on HTTP request and response headers. The framework introduces two new headers: Accept-Datetime and Memento-Datetime. The Accept-Datetime header is used to ask for the version of the original resource valid on that time. If a user agent requests an original resource for a specific datetime, the server responds with a link to the timegate, which can do the datetime content negotiation for that original resource. The timegate redirects the user agent to the appropriate memento, which responds with a memento-datetime. This memento-datetime gives the datetime the resource was created. This datetime of a memento is retrieved using the provenance information of that memento. The provenance of every memento is modeled as a PREMIS OWL Object instance relating to Event instances. Such an Object instance has always a creation event. The datetime of this creation event is used for the Memento datetime content negotiation. List 5 gives an example of such an HTTP interaction.

[24] `http://datatracker.ietf.org/doc/draft-vandesompel-memento/`

```
1: UA ---- HTTP GET/HEAD; Accept-Datetime: Tj ----> URI-R
2: UA <--- HTTP 200; Link: URI-G ------------------- URI-R
3: UA ---- HTTP GET/HEAD; Accept-Datetime: Tj ----> URI-G
4: UA <--- HTTP 302; Location: URI-Mj; Vary; Link:
     URI-R, URI-T, URI-Mj, ------------------------- URI-G
5: UA ---- HTTP GET URI-Mj; Accept-Datetime: Tj ---> URI-Mj
6: UA <--- HTTP 200; Memento-Datetime: Tj; Link:
     URI-R, URI-T, URI-G, URI-Mj ------------------- URI-Mj
```

Listing 5: Typical Memento HTTP interaction

Next to the two new headers, Memento also introduces some new values for the existing HTTP headers: Vary and Link. The value for the VARY header in our case will be *negotiate, accept-datetime, accept*. This VARY header informs that the content negotiation was performed in two dimensions, i.e., the datetime content negotiation and the media type content negotiation. The relation types for the Link header Memento introduced are *original*, for referencing the original resource, *timegate*, for indicating the timegate, *timemap* for linking to the timemap, and *memento* for referencing to various mementos for an original resource. These Link headers allow automatic discovery of the timegate, the timemap, the original resource and several other mementos.

Introducing this Memento datetime content negotiation is justified from our digital long-term preservation perspective. A problem we were facing publishing information as Linked Open Data and preserving it at the same time, involved the enrichments. These enrichments do not always remain valid over time. That is why these enrichments are mostly left out of the metadata to be stored for the long term. If the data providers of the enrichments also support the datetime content negotiation, a memento with enrichments would reference that memento of the enrichment when it was valid. In other words, the Memento datetime content negotiation also preserves the temporality of the information. This justifies storing also the enrichments of the metadata records for the long-term.

5.2 Publishing Provenance

In our platform, every version (memento) of a harvested resource (original resource) has a PREMIS OWL Object description. This Object description describes the provenance of that object and is related through events to object descriptions of other versions/mementos of that original resource. This allows our platform to include in the response of the request for a memento a *provenance* link header which includes the link to the LOD published PREMIS OWL Object description (*URI-Pj*) of that memento. This *provenance* link header will allow automatic discovery of the provenance information.

We extended the Memento framework with a new concept:

- Provenance (URI-Pj): This resource refers to the provenance of the selected version/memento of the original resource.

To allow this resource to be automatically discovered, we extended the Memento framework with a special value for the existing HTTP header Link referencing the provenance information. The relation type for this Link header is *provenance* for the current provenance record (URI-Pj). A typical HTTP interaction, requesting a certain memento, is shown in Listing 6. In our framework steps 1 and 2 of the shown interaction are skipped, because the URI the original resources are published on is also the timegate for the original resources.

The provenance records are themselves also datetime content negotiable. So they become mementos of an original provenance resource. Doing this, gives some extra benefits. The Memento framework defined some extra relation types for the HTTP Link header referencing a memento. When applied to a provenance record of a memento of an original resource, they get the following definitions:

- first memento (URI-M0): This resource refers to the provenance of the first version/memento of the original resource.
- last memento (URI-Mn): This resource refers to the provenance of the last version/memento of the original resource.
- memento (URI-Mj): This resource refers to the provenance of the selected version/memento of the original resource.
- previous memento (URI-Mi): This resource refers to the provenance of the previous version/memento of the selected version/memento of the original resource.
- next memento (URI-Mk): This resource refers to the provenance of the next version/memento of the selected version/memento of the original resource.
- timemap (URI-T): A TimeMap for a provenance record of an original resource lists the URIs of the provenance records of all mementos of that original resource.

The response for a memento request will includes a provenance header link, referencing the provenance information of that memento. This provenance record is on itself also a memento. The response of this memento includes a *timemap* link header pointing to a URI (*URI-T*) listing the URIs of the provenance records of all mementos of that original resource. This way, an agent can have immediately an overall view on the provenance of an original resource.

These extra links could be very helpful in processing the provenance information. Our PREMIS OWL model allows describing digital signatures, signing the versions/mementos of that original resource. A quality checker could investigate the quality and trustworthiness of the published information. This quality checker could investigate the digital signature of the last version. If this was signed by a trusted party and the digital signature is still valid, the quality checker could immediately move on to the provenance of the first memento to check where the signed information

```
1: UA ---- HTTP GET/HEAD; Accept-Datetime: Tj ----> URI-R
2: UA <---- HTTP 200; Link: URI-G ---------------- URI-R
3: UA ---- HTTP GET/HEAD; Accept-Datetime: Tj ----> URI-G
4: UA <---- HTTP 302; Location: URI-Mj; Vary; Link:
     URI-R, URI-T, URI-Mj, ---------------------- URI-G
5: UA ---- HTTP GET URI-Mj; Accept-Datetime: Tj ---> URI-Mj
6: UA <---- HTTP 200; Memento-Datetime: Tj; Link:
     URI-R, URI-T, URI-G, URI-Mj, URI-Pj ---------- URI-Mj
```

Listing 6: Extended Memento HTTP interaction with provenance information

came from and if that data provider is a trusted party also to make a judgment regarding the quality and trustworthiness of the information. The PREMIS OWL model also allows describing the rights information in the provenance of a resource, such as licenses, copyrights, rights granted, etc. A license checker could use these additional links to browse through the provenance records of the mementos of an original resource and check if in none of them violates the rights information of another memento.

A shortcoming of making provenance records also datetime content negotiable, is that all events happening on a preserved resource more recent than the datetime asked for will be left out of the provenance description. Hence, the provenance information would then only contain links to older versions/mementos of the preserved resource and the links to the more recent versions are lost.

To improve the automatic discovery of the provenance information of a memento, our platform will inject the provenance link of the memento also in the HTML and RDF descriptions of that memento. This will enhance the provenance discovery, because not all clients will be able to intercept the *provenance* link header. For the HTML representation of the memento, our framework includes a HTML link tag in the head of the HTML document. This link has a relation type of *provenance*, e.g., <link rel="provenance" href="`http://../object/VTi/1/oai:archipel1.demo.ibbt.be:10_V3`"/>. For the RDF representation, our platform injects a triple denoting the provenance information of that memento. For linking this provenance record (PREMIS OWL Object intance), the PREMIS OWL object property *linkingObject* is used. An example of such an injected triple in the RDF description of a memento is: `<http://../record/VTi/1/oai:archipel1.demo.ibbt.be:10_V3>` premis:linkingObject `<http://../object/VTi/1/oai:archipel1.demo.ibbt.be:10_V3>`.

In some cases, it might be convenient to store the provenance of the provenance information. An example of this in our framework is the characterisation process. This process identifies a memento of an original resource and creates a PREMIS OWL Object instance of it. This can be the metadata record or a multimedia file referenced in a metadata record. In case of a file, the Object description is being enriched with information from the *Preserv2* format registry[25]. This is an enrichment event occurring on provenance information. This could be described in the provenance of the provenance information. Another example of this are digital signatures. Our PREMIS OWL model allows describing these digital signatures applied to a stored memento, but digital signatures can also be used to sign provenance information. When including a *provenance* Link header in

[25] `http://p2-registry.ecs.soton.ac.uk/`

the response to a provenance record, the provenance of the provenance information can be discovered.

Looking at the 5-star deployment scheme[26] of Tim Berners-Lee, this framework could add two more stars for indicating the rating of a Linked Open Data provider. A sixth star could go to Linked Open Data providers that support the Memento datetime content negotiation. This sixth star will indicate to, e.g., a long-term preservation archive, that the enrichments coming from that provider could be stored also for the long term, as discussed earlier. A seventh star could go to Linked Open Data providers not only supporting the Memento datetime content negotiation, but also using this framework to publish their provenance records as Linked Open Data. This seventh star will indicate that the data provider publishes provenance information and, hence, it is possible to make trust judgments over that data using quality checkers or license checkers, as mentioned above.

5.3 Implementation

For implementing this framework, we used Jena TDB as triplestore for the back-end. This is a large-scale persistent triplestore which supports SPARQL. On top of this triplestore, the LOD server was built using Apache Tomcat as HTTP web server. This LOD server has a servlet which will do the datetime and the mediatype content negotiation and will redirect from the original resource, published on a persistent URI, to the appropriate version/memento of that original resource. This servlet will form the timegate. Next to this, we have servlets to serve the appropiate mediatype of the information (HTML and RDF) will also insert the provenance information. The resources that will be published with this timegate are the harvested collections and records. As explained in the previous section, we do not offer datetime content negotiation for the provenance information. For this information, we have a seperate servlet only supporting media type content negotiation.

Next to the LOD server supporting the datetime content negotiation, we have an integration server which will provide the needed preservation processes. These preservation processes will generate the different versions of the harvested information. This integration server was built using the *Porthus .NET* Integration server.

The LOD server will soon be publicly available for demonstration on the URL `http://archipellod.demo.ibbt.be:8080/`. It will support the datetime content negotiation and this can be tested in the *Mozilla* web browser using the Memento plugin[27].

6. CONCLUSIONS

In this article, we have presented a distributed, digital long-term archive relying on semantic technologies. Our platform is able to harvest data, store it for the long-term, and disseminate it as LOD. This data comes from very diverse institutions, each using domain-specific metadata formats. For this, we have developed a layered, semantic metadata model. The top layer lets the archive deal with the diverse data coming from the institutions. For this layer, DC RDF was chosen. The bottom layer will enable the long-term preservation processes and consists of a semantic version of the PREMIS 2.0 data dictionary, i.e., PREMIS OWL. Using this ontology, it is possible to store the metadata needed for the preservation services. It forms the data model for the archive.

A SOA was designed for this distributed archive. This SOA in combination with an ESB allows to modify and expand the current setup of processes and to communicate with all the distributed preservation and dissemination services. This platform produces lots of different versions of the stored information and also produces provenance information, which will relate the different versions of the stored information. To publish these different versions of a preserved resource and their provenance information, our platfrom relies on the Memento datetime content negotiation. We extended this framework to include also HTTP *provenance* header links for automated discovery of the provenance information. This approach allows us to disseminate the versioned information of the preserved resources on persistent URIs, depending on the datetime content negotiation to redirect to the appropriate version/memento of the original stored resource. Combining datetime content negotiation with the publication of the provenance information, links the provenance information to the datetime dimension of a certain stored resource. It also allows to store even the enrichments of the LOD published and preserved resources, because the temporality of these enrichments is also preserved. Finally, the framework allows discovering the provenance information of the other existing versions of an original resource bringing provenance information to the Web. This can all be tested on our publicly available LOD server, published on the following URL: `http://archipellod.demo.ibbt.be:8080/`

7. ACKNOWLEDGMENTS

The research activities that have been described in this paper were funded by Ghent University, K.U. Leuven, VRT-medialab, the Interdisciplinary Institute for Broadband Technology (IBBT) through the Archipel-project (50Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union. Special thanks go out the work package 4 partners. The work described was carried out by this team. This team consists of Gert Goossens, Bart Janssens, and Raf Vandesande from Porthus[28], Descartes, Filip Borloo working for VTi[29], Inge Van Nieuwerburgh and Patrick Hochstenbach from Boekentoren[30], Kris Buytaert from Inuits[31] and Matthias Vandermaesen from Krimson[32].

8. REFERENCES

[1] Bizer, C. and Heath, T. and Idehen, K. and Berners-Lee, T. Linked Data on the Web. In *Proceedings of the 17th International World Wide Web Conference – LDOW Workshop*, pages 1265–1266, Beijing, China, April 2008.

[2] Box, D.; Ehnebuske, D.; Kakivaya, G.; Mayman, A.; Mendelsohn, N.; Frystyk Nielsen, H.; Thatte, S. and

[26] `http://www.w3.org/DesignIssues/LinkedData.html`

[27] `https://addons.mozilla.org/en-US/firefox/addon/mementofox/`

[28] `http://www.porthus.be/default2.aspx`

[29] `http://www.vti.be`

[30] `http://www.boekentoren.be`

[31] `http://www.inuits.be`

[32] `http://www.krimson.be`

Winer, D. Simple Object Access Protocol (SOAP) 1.1, 2000. Available at http://www.w3.org/TR/soap/.

[3] Boyko, A.; Kunze, J.; Littman, J.; Madden, L. and Vargas, B. The BagIt File Packaging Format (V0.96), 2009. Available at https://confluence.ucop.edu/download/attachments/16744580/BagItSpec.pdf?version=1.

[4] Brand, S. Escaping The Digital Dark Age. *Library Journal*, 124, Issue 2:46–49, March 2003.

[5] Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS), Januari 2002. Available at http://public.ccsds.org/publications/archive/650x0b1.pdf.

[6] Gil, Y.; Cheney, J.; Groth, P.; Hartig, O.; Miles, S.; Moreau, L.; da Silva, P. P.; Coppens, S.; Garijo, D.; Gomez, J. M.; Missier, P.; Myers, J.; Sahoo, S.; Zhau, J. Provenance XG Final Report, 2010. Available at http://www.w3.org/2005/Incubator/prov/XGR-prov/.

[7] Hartig, O.; Zhao, J. Publishing and Consuming Provenance Metadata on the Web of Linked Data. In *Proceedings of the 3rd International Provenance and Annotation Workshop IPAW*, 2010. Available at http://olafhartig.de/files/HartigZhao_Provenance_IPAW2010_Preprint.pdf.

[8] Internet Engineering Task Force. RFC 2616: HyperText Transfer Protocol – HTTP/1.1, 1999. Available at http://www.ietf.org/rfc/rfc2616.txt.

[9] Internet Engineering Task Force. RFC 3986: Uniform Resource Identifier (URI) – Generic Syntax, 2005. Available at http://tools.ietf.org/html/rfc3986.

[10] Lagoze, C. and Van de Sompel ,H. The open archives initiative protocol for metadata harvesting - version 2.0, 2002. Available at http://www.openarchives.org/OAI/openarchivesprotocol.html.

[11] Messina, A.; Boch, L.; Dimino, G.; Bailer, W.; Schallauer, P.; Allasia, W.; Basili, R. Creating rich Metadata in the TV Broadcast Archives Environment: the PrestoSpace project. In *IEEE AXMEDIS06 Conference Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution*, pages 193–200, 2006.

[12] Nilsson, M. and Powell, A. and Johnston, P. and Naeve, A. Expressing Dublin Core metadata using the Resource Description Framework (RDF), 2007. Available at http://dublincore.org/documents/dc-rdf/.

[13] Van de Sompel, H; Sanderson, R.; Nelson, M.L.; Balakireva, L.; Shankar, H. and Ainsworth, S. Memento: Time Travel for the Web. *CoRR*, abs/0911.1112, 2009.

[14] Van de Sompel, H; Sanderson, R.; Nelson, M.L.; Balakireva, L.; Shankar, H. and Ainsworth, S. An HTTP-Based Versioning Mechanism for Linked Data. *CoRR*, abs/1003.3661, 2010.