Interweaving OAI-PMH Data Sources with the Linked Data Cloud

Bernhard Haslhofer*
Department of Distributed and Multimedia Systems,
University of Vienna, Vienna, Austria
E-mail: bernhard.haslhofer@univie.ac.at

Bernhard Schandl
Department of Distributed and Multimedia Systems,
University of Vienna, Vienna, Austria
E-mail: bernhard.schandl@univie.ac.at
*Corresponding author

Abstract:
The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has found wide-spread adoption for exchanging bibliographic metadata. In parallel, the W3C's Linked Data Initiative exposes and interlinks structured data from a variety of data sources on the Web. Since many of these data sources contain valuable information for institutional repositories (e.g., shared concept definitions, thesauri, etc.) we believe that institutions that currently expose their data via OAI-PMH can benefit if they integrate their metadata with the data available in the Linked Data cloud. To achieve such an integration, we must bridge the OAI-PMH specific protocol characteristics that currently prevent OAI-PMH metadata from being interoperable with the Linked Data approach of exposing data. As first contribution of this paper, we describe a possible solution for exposing OAI-PMH metadata on the Web as part of the Linked Data cloud. As a second contribution, we present a rule-based mechanism for linking these metadata with other relevant data sources together with a case study that describes possible linking scenarios for three representative OAI-PMH data providers. Finally, we discuss certain quality criteria that OAI-PMH metadata must meet in order to benefit from data exposed by other Linked Data sources.

Keywords: metadata; semantic web; linked data; OAI-PMH; protocols.

Reference to this paper should be made as follows: Haslhofer, B., Schandl, B. (2009) 'Interweaving OAI-PMH Data Sources with the Linked Data Cloud', Int. J. Metadata, Semantics and Ontologies, Vol. X, Nos. 1/2/3, pp.XX–YY.

1 Introduction

The Open Archives Protocol for Metadata Harvesting (OAI-PMH) (Lagoze and van de Sompel, 2002) is currently implemented by more than 1,700 digital library repositories world-wide and enables the exchange of metadata via HTTP. Using this protocol, an application can contact any OAI-PMH data provider and request the bibliographic metadata description of a certain digital or non-digital item. It can, for instance, retrieve metadata describing a historical book archived in the Library of Congress and further process the information that is expressed in the returned metadata description, such as the title, the subject, and the geographic locations described by the book.

In parallel, the W3C's Linking Open Data (LOD) Initiative[1] has started to expose structured data originating from public-interest sources on the Web. The most prominent example of a linked, open data source is DBpedia (Auer et al., 2007), the structured version of the well-known online encyclopaedia Wikipedia, which provides a collection of the world's knowledge from a variety of domains. One of the main characteristics of this initiative is its focus on *linking*, so that exposed datasets contain references to other related and exposed resources. An image published on Flickr, for instance, can be linked to a

[1] `http://linkeddata.org`

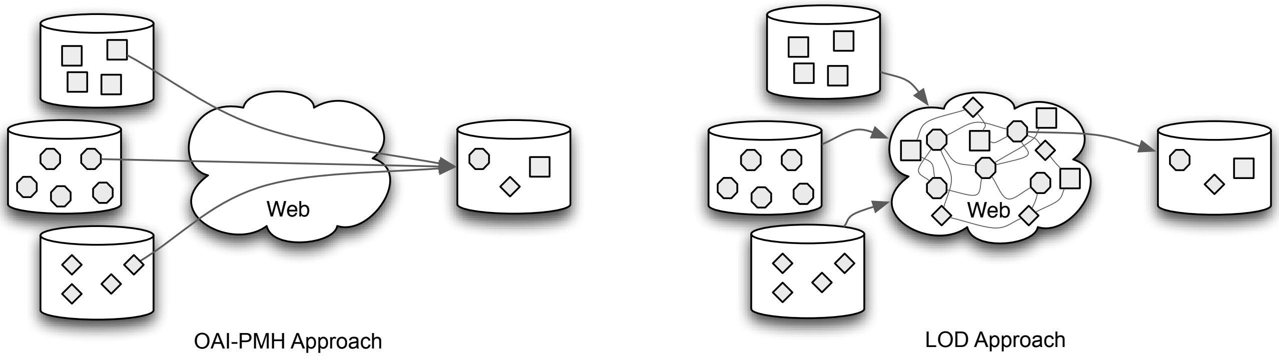


Figure 1: Conceptual differences between OAI-PMH and LOD: OAI-PMH uses the Web infrastructure to transport item descriptions; LOD considers items as parts of the Web itself.

related DBpedia resource representing a Wikipedia article that provides a textual description as well as GIS coordinates of the location where the image was taken.

Interweaving OAI-PMH data sources with the continuously evolving Linked Data can bring the following benefits:

- Metadata that can currently only be harvested by OAI-PMH clients become accessible for a various clients and can easily be integrated in various application scenarios.
- Semantic search engines and crawlers such as Sindice (Oren et al., 2008) can index exposed metadata, which in turn increases the visibility of the contents they describe and the institution that provides the data.
- Clients can follow links to other datasets and combine information that would not be related otherwise.

The current version of OAI-PMH does not allow for such a direct integration. Although it uses Web technologies—in particular HTTP, XML, and URIs—for exchanging metadata, these have mainly the role of a transport layer between repositories. LOD, in contrast, follows the idea of the Web as being "*an information space in which the items of interest (resources) are identified by global identifiers (URI) and which allows embedded references to other URIs*" (Jacobs and Walsh, 2004). Figure 1 illustrates and explains the conceptual differences between the OAI-PMH and the LOD approach.

As a consequence, LOD requires that metadata are not only *exchanged via* the Web but *exposed on* the Web so that each described digital or non-digital item is accessible by a unique dereferencable URI, independent of any OAI-PMH specifics. In a second step, the exposed metadata originating from OAI-PMH data sources must be linked with related data from other sources so that applications can combine these different datasets.

The work reported in this paper, which is an extended and improved version of our previous work (Haslhofer and Schandl, 2008), makes three important contributions: first, after discussing relevant background w.r.t. OAI-PMH and LOD (Section 2), we propose the OAI2LOD Server (Section 3) as a solution that wraps existing OAI-PMH data sources and exposes the available items and metadata *on* the Web according to the Linked Data requirements. Second, we propose a simple rule-based linking approach for enriching the exposed metadata with links to resources in other related data sources based on string similarity metrics. Finally we present the results of the experimental study we have carried out on a representative set of OAI-PMH datasets in Section 4.

2 Background

In this section, we briefly introduce the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Linking Open Data project. Thereafter, we analyse the conceptual differences between these two approaches.

2.1 The OAI-PMH Protocol

The Open Archives Initiatives Protocol for Metadata Harvesting (OAI-PMH) has its origins in the e-print community, where a need for a low-barrier interoperability solution to access distributed and heterogeneous repositories has been identified. Today the protocol is widely used in the digital libraries domain for the exchange and sharing of metadata among repositories. Many popular digital library systems such as the open-source systems Fedora[2], DSpace[3], and EPrints[4] implement the OAI-PMH protocol by default. Client applications can use it to harvest metadata from data providers using open standards such as URI, HTTP, and XML.

2.1.1 Technical Details

The main conceptual entities in the OAI-PMH specification are `Item`, `Record`, and `MetadataFormat`. An item

[2] http://www.fedora.info
[3] http://www.dspace.org
[4] http://www.eprints.org

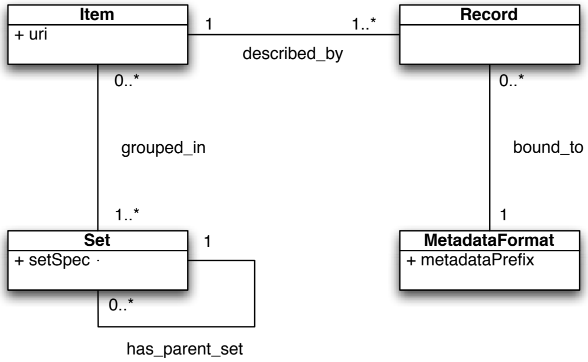


Figure 2: The main conceptual entities of OAI-PMH.

represents a digital or non-digital resource and is uniquely identified by a URI. It can be described by an arbitrary number of metadata records, each of which is bound to a certain metadata format. A record is identified by the combination of the described item's URI and a `metadataPrefix` identifying the metadata format of the record (e.g., `"oai_dc"` for Dublin Core). OAI-PMH further provides the concept of a `Set` for grouping related items and their associated metadata. A set is identified by a `setSpec` parameter (e.g., `"setA"`) and may be part of a set hierarchy, which is indicated by a colon-separated list in the setSpec parameter. The sequence `"setA:setB"`, for instance, means that `setB` is a subset of `setA`. The class diagram in Figure 2 summarises these basic OAI-PMH concepts and their relationships.

OAI-PMH is implemented on top of HTTP. Each OAI-PMH request is actually an HTTP GET request, which contains one of the following *verbs* that specify what kind of information is to be retrieved:

- `Identify` retrieves administrative metadata (e.g., name, owner) about a repository.
- `GetRecord` is used to fetch an individual record for a certain item in a given format.
- `ListRecords` is batch request that harvests all metadata for the entire set of available items in a defined metadata format.
- `ListIdentifiers` returns the identifiers (URIs) of all available items.
- `ListMetadataFormats` enumerates the formats in which the data provider exposes metadata records.
- `ListSets` returns the sets that are available in an OAI-PMH repository.

```
GET /cgi-bin/oai2_0?verb=GetRecord&
    metadataPrefix=oai_dc&identifier=oai:
    lcoa1.loc.gov:loc.gdc/gcfr.0101 HTTP
    /1.1
Host: memory.loc.gov
Accept: */*
```

Figure 3: Sample OAI-PMH request.

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/">
<GetRecord>
...
<record>

<header>
  <identifier>
    oai:lcoa1.loc.gov:loc.gdc/gcfr.0101</identifier>
  <setSpec>ascfrbib</setSpec>
  ...
</header>

<metadata>
  <oai_dc:dc ...>
    <dc:title>Voyage dans les solitudes américaines.
      Voyage au Minnesota,</dc:title>
    <dc:creator>Domenech, Emmanuel, 1826-1903.</dc:creator>
    <dc:subject>Indians of North America.</dc:subject>
    <dc:subject>Minnesota--Description and travel.</dc:subject>
    <dc:description>The principal part of the book is devoted
      to a general account of the Indians of North America,
      following a brief description of Minnesota.
    </dc:description>
    <dc:identifier>
      http://hdl.loc.gov/loc.gdc/gcfr.0101</dc:identifier>
    <dc:language>fre</dc:language>
    <dc:coverage>Minnesota</dc:coverage>
    ...
  </oai_dc:dc>
</metadata>

</record>
</GetRecord>
</OAI-PMH>
```

Figure 4: Sample OAI-PMH response.

Figure 3 shows a sample `GetRecord` request for a Dublin Core metadata record available in the Library of Congress and Figure 4 the received response. The request URI contains the address of the repository, the `GetRecord` verb, the required parameters such as the item's URI identifier (`oai:lcoa1.loc.gov:loc.gdc/gcfr.0101`), and the metadata prefix identifying the desired metadata format (`oai_dc`). The response to an OAI-PMH request is always an XML document. In our example, it consists of a `<header>` section, which contains the item's URI, and a `<metadata>` section encapsulating the metadata record.

2.1.2 Usage of OAI-PMH

There exist a number of OAI Data Provider Registries[5] from which we know that currently at least 1765 institutions worldwide maintain OAI-PMH repositories. Regarding their application domain, we can observe that the protocol has been implemented in a variety of institutions, ranging from small research facilities to national libraries that have integrated this protocol with their catalogue systems. Examples are the *Institute of Biology of the Southern Seas*[6], exposing 403 records, and the *U.S. National Library of Medicine's Digital Archive*[7], exposing 1,272,585 records.

In order to estimate the amount of metadata one can re-

[5] e.g., `http://www.openarchives.org/Register/BrowseSites` or `http://gita.grainger.uiuc.edu/registry/`

[6] `http://repository.ibss.org.ua/dspace/`

[7] `http://www.pubmedcentral.nih.gov/`

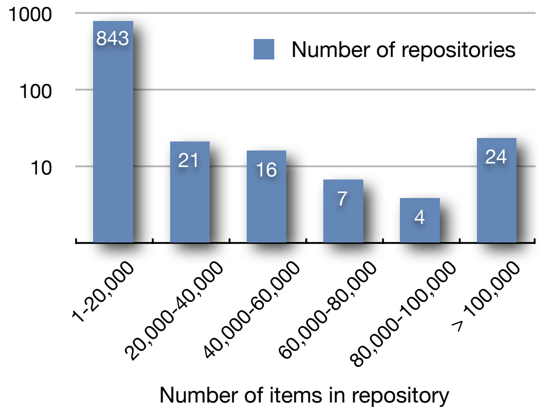


Figure 5: Size of OAI-PMH repositories.

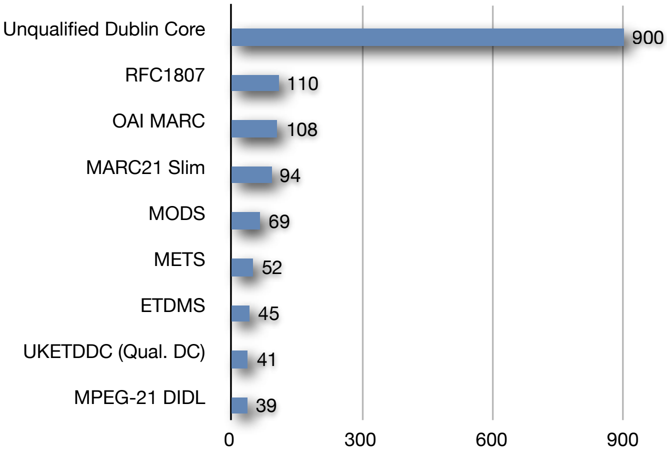


Figure 6: Top 10 Metadata Standards.

trieve via OAI-PMH, we have carried out an analysis on the 915 registered repositories that delivered valid responses. Figure 5 illustrates the size of these repositories using a logarithmic scale on the Y-axis. The results show that 843 (or 92%) of all repositories expose metadata for less than 20,000 items. With 14,303 being the average number of items, the total number of 13,087,842 items is made up of a large number of smaller OAI-PMH repositories.

We expect the number of institutions that expose metadata via OAI-PMH to grow even further. Major attempts of building union catalogues (e.g., the *The European Library (TEL)* project[8]) rely on this protocol for indexing metadata originating from remote sources. Currently, this initiative integrates 47 national libraries and gives access to approximately 150 millions of metadata records. Since the OAI-PMH endpoints of these libraries are currently not listed in the OAI Data Providers Registries mentioned before, we could not consider them in our analysis.

2.1.3 Applied Metadata Formats

In principle, OAI-PMH data providers have the freedom to expose metadata in any format. In order to guarantee a minimum level of interoperability, all data providers *must* at least support the unqualified Dublin Core format (DC, 2006). It is however also recommended to expose metadata in other, semantically richer formats than Dublin Core.

The 915 repositories we have analysed expose metadata in 161 different formats. Besides the mandatory unqualified Dublin Core format[9], RFC1807 (12%), MARC (11.8%) and MARC-21 (10.3%), MODS (7.5%), and METS (5.7%) are the most frequently used formats[10]. The large gap between Dublin Core and the other metadata formats reveals that most data providers do not follow the OAI-PMH standard's suggestion of exposing metadata in a semantically richer format rather than unqualified Dublin Core. Figure 6 summarises these results.

2.2 The Linking Open Data Project

Linked Open Data (LOD) denotes a specific style of information publishing on the Web. This style has been proposed by a W3C community project whose main goal is to increase the value and applicability of public information by (a) exposing them using standardised formats and protocols, (b) expressing them using defined, shared vocabularies, and (c) interlinking information from different sources in order to provide a global, machine-interpretable information graph.

The value of Linked Open Data is made obvious by comparing it with the World Wide Web, which consists of numerous web pages that are intended to be readable for humans: they are linked so that a user can navigate from one page to another. The World Wide Web works because there exist standardised identifier formats and protocols (URI/URL, HTTP), a globally accepted vocabulary for information representation (HTML), and hyperlinks between pages. With one of these components missing, the Web would not be as usable and useful as it is.

In Linked Open Data, these building blocks are applied to the domain of publicly available, machine-interpretable data. LOD emerged from the observation that many institutions, which offer their data to the public, do neither provide unified access mechanisms nor links to other relevant data sources. Because of these conditions, many institutions' datasets remain to be isolated and require specialised client applications to make use of the data. A generic LOD client, however, is able to retrieve and integrate data from arbitrary institutions, which has significant advantages both for the client and for the data provider.

2.2.1 LOD Principles

LOD requires a data publisher to follow four simple rules, as defined by Berners-Lee (2006) and further elaborated by Bizer et al. (2007). These rules ensure that the main building blocks of LOD, as described above, are maintained. In the following we give a short discussion of these four rules.

[8] `http://www.theeuropeanlibrary.org`

[9] 15 of the 915 repositories do not implement the mandatory Dublin Core format and are therefore not OAI-PMH compliant.

[10] Further information about these standards: `http://www.loc.gov/standards` and `http://rfc.net/rfc1807.html`

1. *Use URIs as names for things.* LOD heavily relies on the concept of URIs (Berners-Lee et al., 2005) for the identification of "things", i.e., objects or resources that are described by the exposed data. To facilitate interlinking between different datasets, it is important to define a permanent, globally unique URI for each item under consideration.

2. *Use HTTP URIs so that people can look up these names.* LOD proposes to use URIs not only for identification, but also for physical access to information representation. The HTTP protocol (Fielding et al., 1999) is a well-established means for transporting data from a server to a client, and it provides several features that make it well suitable for the needs of LOD. One important feature is *content negotiation*[11], which makes it possible to serve different representations of a resource under the same URI. Applications requesting a resource can specify the representation to be returned by using the HTTP `Accept` header field. Hence it is recommended to use dereferenceable HTTP URIs as identifiers, and hence make LOD resources part of the World Wide Web.

3. *When someone looks up a URI, provide useful information.* While this rule may seem obvious at the first sight, it turns out that it is not straightforward to define what *useful* means. On the Web, usefulness refers to both the syntactical as well as the semantic layer of information: a HTML page in Latin language may be useless to somebody not aware of Latin, although it will be properly rendered by the browser. In terms of syntax, the common representation format is RDF (Klyne and Carroll, 2004); and it is considered good practice to use terms from existing vocabularies and ontologies (cf. Bizer et al. (2007), Section 4). It is, however, also recommended to provide a human-readable representation of items so that a client that retrieves data can select the appropriate representation through HTTP content negotiation (cf. Bizer et al. (2007), Section 2.1). Again, content negotiation allows the client to specify which formats it is able to interpret.

4. *Include links to other URIs, so that clients can discover more things.* The true power of Linked Open Data lies in the possibility to define links between resources, which is a fundamental principle in Web-based design. When a data provider publishes information as LOD, not only internal links (i.e., links within the published dataset), but also external links (links that refer to items from other datasets) should be set where appropriate. Similar to the World Wide Web, this ensures that a client can make use not only of one publisher's data, but is also enabled to follow the links and combine data from different sources.

Additionally it is recommended for LOD data providers to offer a SPARQL endpoint (Prud'hommeaux and Seaborne, 2008; Clark et al., 2008). Using this RDF query language, clients can access Linked Data in a more selective manner and retrieve exactly the data they need. For instance, SPARQL can be used for selective batch retrieval: it is possible to ask for item descriptions based on metadata criteria, to limit the number of results, or to transform data according to the client's needs.

With these four rules in place, information providers can open their datasets to the wide public and provide the premises to participate in a global network of interlinked information, the so-called *Web of Data*[12]. This, in consequence, enables them to actively benefit from a large number of applications, some of which we describe in the following section.

2.2.2 Possible Applications

The range of applications that can benefit from data published as LOD is as wide as the range of applications that the World Wide Web provides. One possible application field is *information search and retrieval.* There already exist a number of search engines for Linked Data; one of the most prominent ones is Sindice (Oren et al., 2008). While current Web search engines only provide search results based on full-text indices and structural metrics (like the number of links that refer to a Web page), a semantic search engine operates on structured, machine-interpretable data and hence is able to also take into account the applied vocabularies and ontologies, in order to provide more fine-grained and expressive search facilities.

The fact that LOD uses structured vocabularies to describe data is also utilised by a number of generic LOD *data browsers*; e.g., Tabulator (Berners-Lee, 2006). In contrast to web browsers, whose functionalities are limited to the pure rendering of content, LOD browsers can make use of the data-inherent structures and, if well-known vocabularies are used to describe resources, of their semantics, in order to improve information rendering. For instance, the Humboldt LOD browser allows users to explore data in a way that is customised to the current situation and context by the combination of browsing with faceted filtering (Kobilarov and Dickinson, 2008).

Of course, Linked Open Data can be viewed not only through generic browsers: the usage of application- and provider-independent vocabularies allows LOD data to be integrated into existing portals and applications. For instance, in the cultural heritage domain it is common for institutions to offer access to their datasets via Web portals that visually represent the described items. If an institution applies the LOD guidelines to its data and interlinks them with external data sources, the portal can be dynamically enriched with additional information.

[11] For a more detailed explanation we refer to Section 12 of the HTTP/1.1 specification (Fielding et al., 1999).

[12] `http://www.w3.org/2001/sw`

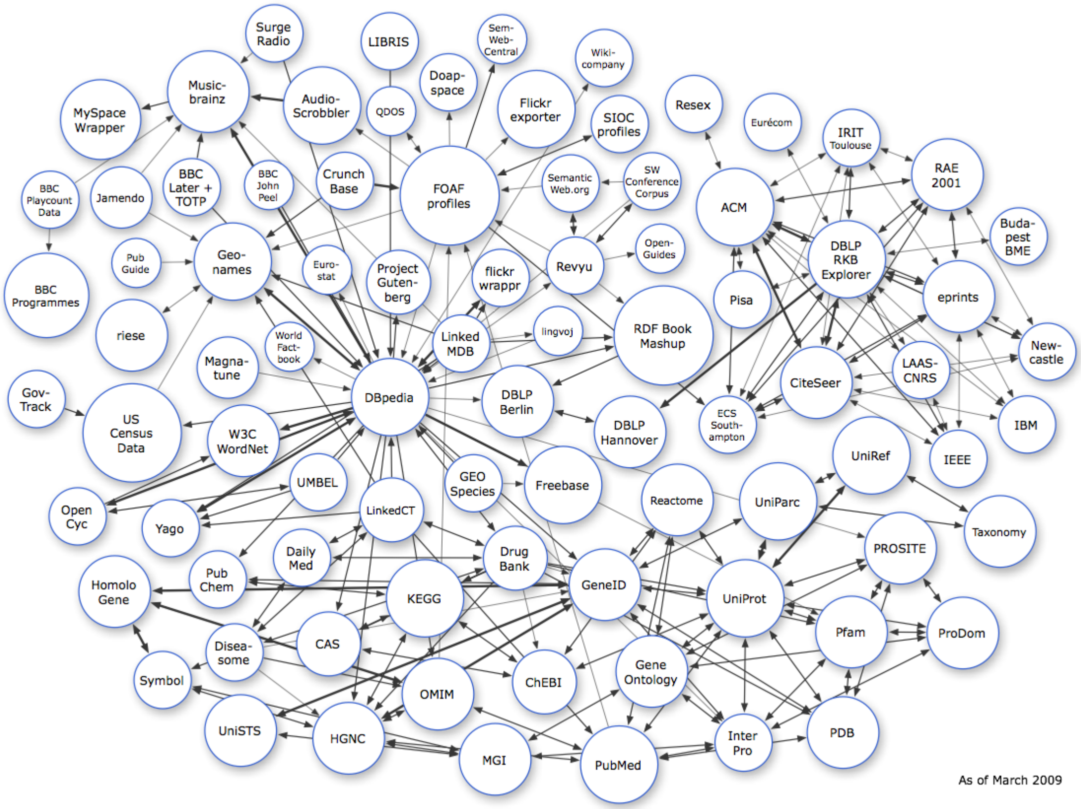


Figure 7: The Linked Open Data Cloud (Jentzsch, 2009).

2.2.3 Available LOD Data Sources

A fair number of datasets are already available in the form of Linked Open Data (an overview is given in Figure 7), and their number is growing continuously[13]. One of the most important datasets is DBpedia (Auer et al., 2007), a LOD representation of Wikipedia articles. The DBpedia service defines a unique URI for each Wikipedia article, converts semi-structured information (e.g., from Wikipedia info-boxes) into a structured RDF representation, and allows clients to access and query descriptions. Since it is not restricted to a specific domain, DBpedia data can be of use for many different applications, and because of its high degree of interlinkage it acts as a hub for different datasets. Other relevant datasets, which are relevant for the digital libraries and archives domain, include the recently published set of Library of Congress Subject Headings (LCSH) (Summers et al., 2008) and the Swedish Union Catalogue (Malmsten, 2008).

LOD does not define how published data is persisted. For data that is already available in a (semi-)structured format, a number of *wrappers* have been proposed, which are able to expose data as LOD with very limited configuration overhead. A number of such wrappers are already available, e.g., D2RQ for relational data bases (Bizer and Seaborne, 2004), or OpenLink Virtuoso Sponger[14] which is an extensible framework for the transformation of a variety of data formats to RDF. In Section 3 we present the OAI2LOD server, a wrapper component that is able to expose OAI-PMH endpoints as Linked Data.

2.3 The Conceptual Gap between OAI-PMH and LOD

OAI-PMH has been designed for transferring large amounts of metadata from a server to a client *via* the Web. From that perspective, it provides a reasonable solution for clients that need to aggregate or index metadata from remote repositories. The goal of the LOD initiative, however, is a different one: it aims at exposing metadata *on* the Web as machine and human-readable data that describe certain resources, which in turn are identified via dereferencable URIs. An additional goal is provide structured query access to these data via SPARQL, which is both an RDF query language and a Web-based query protocol.

We have identified the following conceptual differences between OAI-PMH and LOD:

- Both protocols use URIs for the identification of resources (items in the case of OAI-PMH); in OAI-PMH, however, these URIs serve solely for identification purposes whereas in LOD, URIs take the role of dereferencable identities.

[13] LOD Datasets: `http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets`

[14] `http://docs.openlinksw.com/virtuoso/virtuososponger.html`

- OAI-PMH introduces protocol specific verbs (e.g., GetRecord) and a set of adjacent parameters. Clients must be aware of these verbs and parameters in order to be able to retrieve metadata from a remote repository. LOD, in contrast, builds on the functionality provided by existing Web technologies; e.g., standard HTTP methods. As a result, LOD data are accessible for any client that is aware of HTTP, URI, and HTML or RDF, respectively.

- LOD relies on the built-in HTTP content negotiation features in order to deliver data in various representations; OAI-PMH is restricted to XML as the only valid representation format.

- OAI-PMH provides batch retrieval functionality, which enables the transfer of a large amount of metadata descriptions within a single HTTP transaction. In LOD, such functionality is provided indirectly by the SPARQL query language and protocol. SPARQL allows the formulation of complex selection criteria and provides LIMIT and OFFSET clauses to return metadata that match certain criteria (e.g., "all records describing items created by X") or even just a subset of the available metadata values (e.g., "all authors of all books in a library").

- OAI-PMH can return metadata records for one and the same item in several metadata formats. When following the LOD design principles without tailoring them to OAI-PMH specific needs, one cannot request specific metadata formats from a LOD endpoint. It is however possible to describe a resource with different vocabularies and to use the SPARQL query language to return metadata in certain formats only.

- OAI-PMH supports a kind of version control and allows clients to retrieve only those metadata records that were created or modified in a given date-range, specified by from and until attributes in the ListRecords and ListIdentifiers requests. One possible approach in the context of LOD is to keep so called *linked data update logs*, as described in Auer et al. (2009). Another simple, straightforward solution is to introduce OAI-PMH specific vocabulary terms and use SPARQL to query for date ranges in order to retrieve the resource that have been created or modified within a specific date range.

Regarding these conceptual differences, we can observe that the LOD approach already subsumes a large fraction of the functionality provided by OAI-PMH, even though in a slightly different way. This implies that if existing OAI-PMH data providers publish their metadata on the Web by following the LOD principles, any client can fetch and process these metadata by simply crawling and resolving the exposed, dereferencable URIs in a certain URI domain space.

3 The OAI2LOD Server

As a possible solution for bridging the conceptual differences between OAI-PMH and LOD we propose the *OAI2LOD Server*, a wrapper component that can expose OAI-PMH compliant data sources as Linked Data on the Web. It allows institutions to interweave their metadata with the Linked Data cloud by instantiating the OAI2LOD Server as a gateway between their existing OAI-PMH endpoint and the Web of Data. In the following we describe how the OAI2LOD Server bridges the conceptual gap between OAI-PMH protocol specifics and the LOD principles.

3.1 Dereferencable Item and Set Identifiers

According to the first Linked Data rule, things should have URIs (cf. Section 2.2.1). If we regard the OAI-PMH protocol concepts, the entities Item, Set, and MetadataFormat are such things. In Table 1, we depict an example mapping between OAI-PMH identifiers and corresponding LOD URIs. For items, the OAI-PMH specification already demands that each item must be identified by a URI; this is not the case for sets: they are identified by arbitrary strings consisting of any valid unreserved characters, which may form a colon-separated list if sets are arranged in a hierarchy. Such strings, however, are no valid URIs because they do not define a mandatory scheme component (Berners-Lee et al., 2005). Also metadata formats are not directly addressed by their URIs but by their metadata prefixes (e.g., oai_dc). But, since a metadata prefix always resolves to an XML namespace URI that serves as global identifier for a format, we can regard metadata formats as things that are identified by URIs, although these are not necessarily dereferencable HTTP URIs. We further discuss vocabulary related issues in Section 3.2.

The second Linked Data rule demands that URIs that identify things should be resolvable HTTP URIs, i.e., URLs that return some information when being dereferenced with an HTTP request. As mentioned before, OAI-PMH uses non-resolvable URIs, i.e., URNs, to identify items and metadata formats, and non-URI strings to identify sets. For items and sets we can bridge that gap by wrapping the item and set identifiers with resolvable HTTP URLs using the pattern described in Figure 8. Following the generic URI syntax, the element <authority> denotes the naming authority that is in charge of the name space or domain that is defined as part of the URI (e.g., the Library of Congress: memory.loc.gov). The elements <identifier> and <setSpec> refer to the item and set identifiers as defined in the OAI-PMH specification (cf. Section 2.1). The resulting URIs are depicted in Table 1.

Following the third Linked Data rule, one should deliver useful information whenever a URI is dereferenced. Hence, we should deliver metadata records in a format that can be processed by a variety of applications. Currently, the OAI-PMH protocol delivers metadata for clients that are aware of the OAI-PMH protocol specifics, know the under-

Resource Type	OAI-PMH Identifier	LOD URI
Item	oai:lcoa1.loc.gov:loc.gdc/gcfr.0101	http://memory.loc.gov/resources/item/oai:lcoa1.loc.gov:loc.gdc/gcfr.0101
Set	ascfrbib	http://memory.loc.gov/resources/set/ascfrbib
Item Metadata Record	n/a	http://memory.loc.gov/rdf/item/oai:lcoa1.loc.gov:loc.gdc/gcfr.0101
Set Metadata Record	n/a	http://memory.loc.gov/xhtml/set/ascfrbib
Metadata Format	http://www.openarchives.org/OAI/2.0/oai_dc/	http://purl.org/dc/elements/1.1/

Table 1: Example mappings of OAI-PMH Identifiers to LOD URIs.

```
http://<authority>/resource/item/
    <identifier>
```

```
http://<authority>/resource/set/<setSpec>
```

Figure 8: Wrapping pattern for item and set identifiers.

```
http://<authority>/<format>/<oai-pmh-
    concept>/<identifier>
```

Figure 9: Pattern for naming various resource representations.

lying XML Schema definition[15], and are able to parse the returned XML responses; so it returns metadata that is useful for OAI-PMH clients only. In order to deliver useful information also for other clients, such as Web browsers or Web crawlers that build up the indexes for (semantic) search engines, the OAI2LOD Server exposes information in various representations using content negotiation as explained in Section 2.2.1. At the moment, XHTML, JSON, and RDF serialisation formats, i.e., RDF/XML and N3, are supported. While Web browsers can process the former and display the exposed information to humans, the latter formats can be processed by machines. Since for LOD it is recommended that also the various representations of a resource have unique URI identifiers, we propose the pattern described in Figure 9 for naming the various representations. The element `<format>` refers to representation format (e.g., XHTML or RDF), and the element `<oai-pmh-concept>` denotes the exposed OAI-PMH concept, i.e., `item` or `set`.

3.2 Vocabularies

The first three LOD rules also apply to vocabularies and demand that the definitions of the metadata elements that are returned as part of an HTTP response (e.g., `dc:creator`) are dereferencable URI resources. We can achieve that by exposing metadata vocabularies in various representations on the Web, in a similar manner as this has been done for the Dublin Core definition (e.g., `http://purl.org/dc/elements/1.1/creator`). Again, HTTP content negotiation allows clients to retrieve a schema definition either in XHTML or RDF Schema. If institutions would like to expose their metadata in different formats than Dublin Core, they must assure that these formats are expressed in RDFS and/or OWL and published as dereferencable vocabularies on the Web. Detailed instructions on how to publish vocabularies on the Web have been discussed by Berrueta and Phipps (2008).

For resembling OAI-PMH semantics when exposing metadata on the Web, we have introduced an OAI-PMH specific vocabulary that defines the main protocol concepts in RDFS and expose them as dereferencable URIs on the Web. In particular, we have currently published the vocabulary within the dereferencable namespace `http://www.mediaspaces.info/vocab/oai-pmh.rdf`. The main concepts defined therein are the classes `Item` and `Set`, as well as properties like `setName` and `setDescription`.

3.3 Linking OAI-PMH with LOD Sources

The fourth Linked Data rule proposes to include links to other URIs, so that machines can discover more things. In the context of the OAI2LOD Server this means that the metadata records describing the exposed items should contain links to resources in other, related data sources. Example linking possibilities include

- links between equal items that are described by different providers (e.g., a book may be present in several libraries); and

[15] OAI-PMH schema: `http://www.openarchives.org/OAI/2.0/oai_dc.xsd`

- links to items in external LOD sources such as Wikipedia, the Library of Congress Subject Headings (LCSH), or any other data source that provides additional, relevant information about a certain item (cf. Section 2.2.2). A Wikipedia article, for instance, could provide information such as an author's biography, a link to an image depicting the author, and many other information that is not explicitly available in an item's metadata record. A reference to a LCSH term could be exploited in order to take broader and narrower terms into account, which can be useful for search and discovery purposes.

A precondition for enriching metadata records with links is that the related data sources expose their data according to the LOD principles as well and that they provide a SPARQL query interface. The OAI2LOD Server provides an automated linking job that compares the resource descriptions provided by a certain OAI-PMH endpoint (source records) with resource descriptions in a defined remote data source (target records). The linking job is freely configurable and takes a set of linking rules as input. A linking rule defines which metadata elements in the source records should be compared with which elements in the target records. The linking job translates the linking rule into a SPARQL query, uses that query to fetch the required data from the specified remote data source, and compares the elements (properties) and their lexical values with each other. For each linking rule, the administrator can decide which string comparison heuristics to apply and define a similarity threshold $0 \leq \epsilon \leq 1$ above which two strings are considered to match each other. Also, the administrator can define what type of link (`linkingProperty`) should be added to a source record if there is match with a certain target record. This of course depends on the semantics of the elements that are considered in the linking job. Possible linking properties that are already provided by common Semantic Web ontology languages are

- `rdfs:seeAlso`, which specifies a resource that might provide additional information about the subject resource; and

- `owl:sameAs`, which indicates that two URI references actually refer to the same real-world thing.

Figure 10 shows an excerpt of an OAI2LOD Server configuration file, which defines a single mapping rule for creating links with resources defined in DBpedia. In particular, the rule compares the Dublin Core property `dc:coverage` of all `oai:items` in the OAI2LOD Server instance with all `rdfs:labels` of all resources in DBpedia that are classified as YAGO category `yago:StatesOfTheUnitedStates`[16]. For comparing the values of the elements in the source and target records, it uses the `Levenshtein`[17] string similarity metrics and adds a link of type `rdfs:seeAlso` to a source record if the similarity threshold $\epsilon \geq 0.98$.

```
<lrule1> a oai2lod:LinkingRule;
    oai2lod:sourceType oai:Item;
    oai2lod:sourceProperty dc:coverage
    oai2lod:targetType yago:
       StatesOfTheUnitedStates;
    oai2lod:targetProperty rdfs:label;
    oai2lod:linkingProperty rdfs:seeAlso;
    oai2lod:similarityMetrics <http://
       dbpedia.org/resource/
       Levenshtein_distance>;
    oai2lod:minSimilarity "0.98"^^xsd:
       float .
```

Figure 10: Sample OAI2LOD linking rule.

Although the OAI2LOD linking mechanism integrates the whole spectrum of string similarity metrics that is part of the SimMetrics library[18] provided by the University of Sheffield, determining matches between distinct resources is a non-trivial task that requires careful administration to avoid the addition of semantically incorrect links. For certain elements (e.g., textual descriptions), comparing the instance values is unsuitable because determining a match between two syntactically different but semantically corresponding descriptions is a hard task for machines. For other fields with a well-defined, precise syntax and semantics (e.g., data elements using a specified encoding schema such as ISO 8601 country codes (ISO TC 154, 2004)) automated linking is possible in a precise and semantically correct way. For domain specific string metrics, the OAI2LOD Server provides the possibility to include domain-tailored string similarity algorithms.

Figure 12 shows the RDF/XML representation of our example metadata record as it is returned by the OAI2LOD Server as a response to the request depicted in Figure 11[19]. It contains the same information as the record in Figure 4 but represents the metadata according to the Linked Data principles. The record also includes a `rdfs:seeAlso` link to a DBpedia resource representing an article about Minnesota, i.e., additional information about the location the described book is about. As a possible application scenario, an RDF-aware application could then follow that link, extract the GIS location from the structured DBpedia resource description, and place the book on a virtual map.

3.4 Batch Harvesting

With its `ListIdentifiers`, `ListRecords`, and `ListSets` verbs, the OAI-PMH protocol provides the necessary functionality for retrieving item identifiers, metadata records, or set descriptions in a batch-manner, i.e., to retrieve,

[16] For further details on the YAGO ontology, we refer to Suchanek et al. (2007).

[17] see (Levenshtein, 1966)

[18] SimMetrics Similarity Metric Library: `http://sourceforge.net/projects/simmetrics/`

[19] Since the domain `memory.loc.gov` is not under the author's control, this is a fictitious example

```
GET /resources/item/oai:lcoa1.loc.gov:loc.
    gdc/gcfr.0101 HTTP/1.1
Host: memory.loc.gov
Accept: application/rdf+xml
```

Figure 11: Sample OAI2LOD Server request.

```
<rdf:RDF
  xmlns:oai="http://www.mediaspaces.info/vocab/oai-pmh.rdf#">
  ...

<oai:Item
  rdf:about="http://memory.loc.gov/resource/item/
  oai:lcoa1.loc.gov:loc.gdc/gcfr.0101">

  <dc:title>Voyage dans les solitudes américaines.
    Voyage au Minnesota,</dc:title>
  <dc:creator>Domenech, Emmanuel, 1826-1903.</dc:creator>
  <dc:subject>Indians of North America.</dc:subject>
  <dc:subject>Minnesota--Description and travel.</dc:subject>
  <dc:description>The principal part of the book is devoted
    to a general account of the Indians of North America,
    following a brief description of Minnesota.
  </dc:description>
  <dc:identifier
    rdf:resource="http://hdl.loc.gov/loc.gdc/gcfr.0101"/>
  <dc:language>fre</dc:language>
  <dc:coverage>Minnesota</dc:coverage>

  <oai:set
    rdf:resource="http://memory.loc.gov/resource/set/ascfrbib"/>

  <rdfs:seeAlso
    rdf:resource="http://dbpedia.org/resource/Minnesota"/>
  ...
</oai:Item>

</rdf:RDF>
```

Figure 12: Sample OAI2LOD Server response, linked to a relevant DBpedia article.

for instance, 100 records by issuing a single HTTP `GET` request. Especially for large datasets and frequent harvesting intervals this feature reduces the overall harvesting time and network load. For partitioning the returned data into chunks of a specified size and controlling the data flow between a data provider and the requesting clients, OAI-PMH introduces the concept of *resumption tokens*, which are returned as part of a batch harvesting response. In order to retrieve the next portion of the complete list, a subsequent request must contain the value of that key. In the example request URL `/?verb=ListRecords&resumptionToken=1234`, the character sequence `1234` represents the resumption token.

Without tailoring the LOD principles to OAI-PMH specifics, the best solution for retrieving metadata from a remote host in a batch manner is to use the SPARQL query language with its `LIMIT` and `OFFSET` patterns. Listing 13 illustrates the SPARQL query pendants to the given OAI-PMH batch retrieval requests: the first query uses a simple `SELECT` statement that returns the first 100 resources of type `oai:Item`, i.e., the first hundred item identifiers. By using a SPARQL `DESCRIBE` query, one can retrieve the metadata records for a given number of items — in this example 50 records starting from position 300. Batch Retrieval of set descriptions works in an analogue way.

```
//Retrieving identifiers in a batch manner

SELECT ?x
    WHERE {?x a oai:Item}
OFFSET 0
LIMIT 100


//Retrieving records in a batch manner

DESCRIBE ?x
    WHERE {?x a oai:Item}
OFFSET 300
LIMIT 50
```

Figure 13: Batch requests in the context of LOD.

```
http://<authority>/<format>/all/<oai-pmh-
   concept>
```

Figure 14: URI names for batch dataset representations.

The advantage of using SPARQL queries for batch retrieval is that one can further restrict queries to a certain subset of the available metadata. Although OAI-PMH supports selective harvesting, the possible restriction dimensions are limited to records that were created, deleted, or modified within a specific date range of records that describe items that belong a certain set. With SPARQL we can further extend these possibilities by introducing additional, application-specific restrictions; we could for instance retrieve only the `dc:creator` elements of those metadata records that were modified after a specific date and describe items that are part of a certain set.

For the batch retrieval of item identifiers the OAI2LOD Server provides an alternative to the SPARQL approach. If a client request issues the request pattern described in Listing 8 without including a specific `<identifier>` in the URI path, the response delivers a list of all item identifiers. Since the OAI2LOD Server can return this list in various representations, depending on the requested format (e.g., RDF, XHTML), we have named these data representations with dereferencable URIs that follow the pattern described in Listing 14.

3.5 Selective Harvesting

Via OAI-PMH one can retrieve metadata records in a selective manner by including the `from` and `until` attributes with appropriate datestamps in `ListIdentifiers` or `ListRecords` requests. As a response the client receives those metadata records (or identifiers) that have been been created or modified within the given date range. Depending on the OAI-PMH repository capabilities the responses may also include the headers of deleted records with a corresponding status attribute. In fact, we can conceive this as basic version control functionality.

Although not implemented in the current version (v.0.2), the OAI2LOD Server could easily provide this functionality by storing the creation and modification dates with each harvested item. This, of course, requires that the OAI-PMH specific vocabulary described in Section 3.2 is extended with the necessary terms in order to reflect that information (e.g., `oai:modificationDate`). For deleted metadata records the OAI2LOD Server could follow the same strategy as OAI-PMH repositories and keep their status (e.g., `oai:status = deleted`). Clients can make use of these selective harvesting features by issuing SPARQL queries that include triple patterns or filters for date ranges or status conditions.

3.6 Implementation

The OAI2LOD Server, as illustrated in Figure 15, is a stand-alone server implemented in Java, and is based on the architecture of the D2RQ Server (Bizer and Seaborne, 2004). It can be configured to expose records from a specific OAI-PMH endpoint in a certain metadata format, according to the principles described above. A scheduled process regularly harvests metadata from the given endpoint, transforms them into RDF/XML using a format-specific XSL style-sheet, stores the transformed metadata in a built-in triple store, and exposes the metadata to clients. The built-in request handler/dispatcher analyses the HTTP `Accept` header contained in a client request and redirects the client to the corresponding entry point, which provides metadata in the appropriate representation, using the `HTTP 303 See Other` response.

4 Case Study

As a proof of concept of our OAI2LOD Server approach, we have carried out a case study and exposed metadata from three representative OAI-PMH data providers as Linked Data on the Web. The goal of the study was to obtain a first impression on the practical applicability of the OAI2LOD Server implementation and to receive qualitative feedback on the provided linking mechanism. We are aware that our linking algorithm does not fully reflect the possibilities of currently known entity resolution or record linkage solutions (cf. Section 5) and therefore, at this place, we refrain from an evaluation on the qualitative and quantitative performance of the linking algorithm in comparison to existing solutions. We rather would like to describe three linking scenarios, which can give potential adopters of the OAI2LOD Server valuable hints on how to define linking rules. For this purpose we have set up three OAI2LOD Server instances for the following data providers and metadata records they expose:

- The *Library of Congress (LOC)* OAI-PMH endpoint exposing Dublin Core metadata.
- The *National Library of Australia (NLA)* OAI-PMH endpoint exposing Dublin Core metadata.
- The *Austrian National Library's Image Archive (ONB)* exposing metadata in a proprietary format.

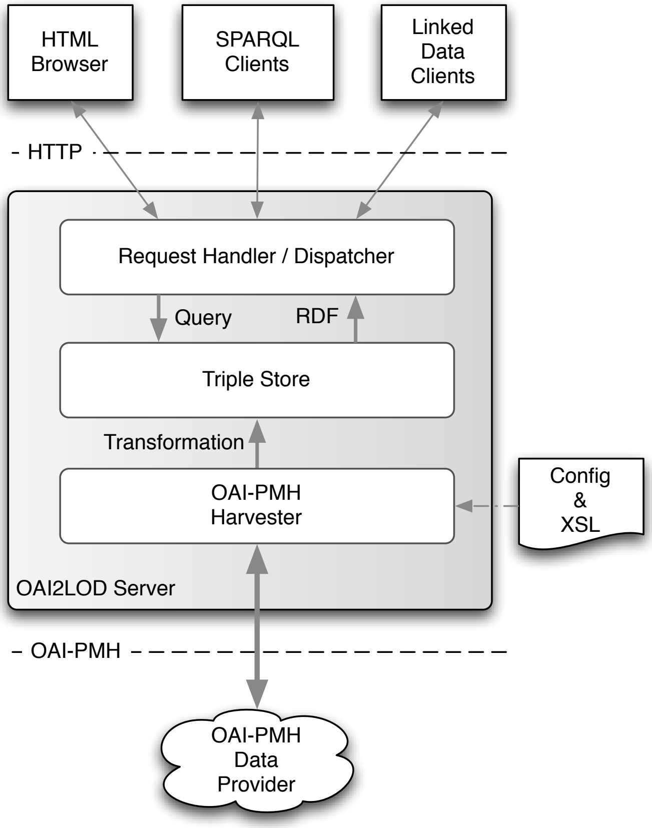


Figure 15: The OAI2LOD Server architecture.

For each of these data providers, we have set up a separate OAI2LOD Server instance that harvests 5000 metadata records and interlinks them, if applicable, with resources in DBpedia according to predefined linking rules. From the generated data we have randomly chosen 100 links and manually checked their semantic validity. The configuration files we used for this case study are part of the OAI2LOD open source distribution[20], which allows a reproduction of the described experiments.

4.1 Dataset 1: Library of Congress — Dublin Core

The Library of Congress OAI-PMH repository, available at `http://memory.loc.gov/cgi-bin/oai2_0`, exposes metadata from selected collections of historical materials, including many from American Memory and the Print and Photographs Online Catalogue. This includes photographs, movies, maps, books, etc. Since the OAI2LOD Server supports the exposition of Dublin Core metadata by default, we only needed to adapt the OAI-PMH endpoint and the number of records to be harvested

[20] OAI2LOD Server: `http://www.mediaspaces.info/tools/oai2lod/`

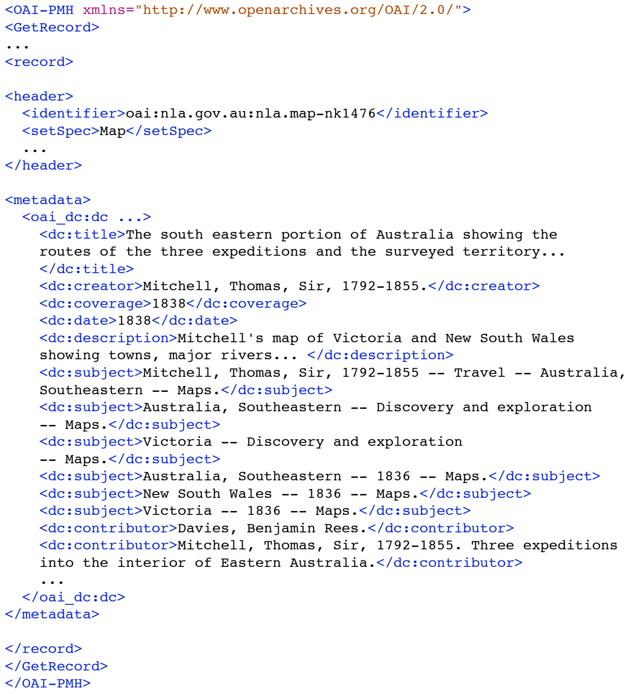
in the configuration file. We applied the linking rule presented in Figure 10, i.e., we compared the values given by the `dc:coverage` field with `rdfs:label` properties of all resources in DBpedia that are classified within the YAGO category `yago:StatesOfTheUnitedStates`. Hence, the goal of the linking mechanism was to link the spatial topics of the LOC metadata with DBpedia articles that further describe the mentioned locations.

Our test set consisted of 5,000 Dublin Core metadata records that were harvested through iterative `ListRecords` requests. In total, the records contained 11,026 `dc:coverage` property-value pairs. With a similarity threshold of 0.98, the linking mechanism created 774 links between LOC and DBpedia resources, which were added to 653 records (94 records were annotated with more than one link to a DBpedia concept). All links were considered correct in a manual examination, i.e., all items covering a certain state in the united states were linked to a DBpedia resource representing an article about the respective state. However, a number of possible links were not created by our algorithm for two reasons: first, a number of states in DBpedia (e.g., `dbpedia:Georgia` or `dbpedia:Florida`) were not categorised as `yago:StatesOfTheUnitedStates` and therefore not recognised as possible linking candidates. The second reason was that the label correspondence between the LOC dataset and DBpedia articles did not fit, either because LOC uses different strings to identify the same state (e.g., `"Virginia"`, which we were able to match to `dbpedia:Virginia`, compared to `"United States--Virginia"`, which we were not able to match), or because the `dc:coverage` field in the LOC dataset did not precisely describe a state of the United States (e.g., `"California, Southern"`).

4.2 Dataset 2: National Library of Australia — Dublin Core

The OAI-PMH handler of the National Library of Australia Digital Object Repository, available at `http://www.nla.gov.au/apps/oaicat/servlet/OAIHandler`, exposes the NLA's digital collections including pictures of people and places in Australia, music collections, manuscripts, as well as books and serials. Currently, it exposes metadata in Dublin Core and a proprietary schema identified by the prefix `do`, which extends the Dublin Core by three technical field such as `do:thumbnail` or `do:mediumresolution`. Figure 16 shows an example metadata record harvested from the NLA repository.

The problem in this case was that the NLA metadata were flattened into a set of Dublin Core elements and therefore not clearly semantically distinguishable. The `dc:subject` element, for instance, can contain a variety of metadata ranging from person names over geographical regions to other concepts such as `Maps` or `Discovery and exploration`. Hence, terms with a broad spectrum of meanings have been combined and flattened into a single element. In such a case, our linking mechanism is not applicable because (i) it is unclear how to restrict the set of possible DBpedia linking candidates to a certain type (i.e., YAGO category), and (ii) a string comparison would result in a very low similarity if strings such as `"Australia, Southeastern -- Discovery and exploration -- Maps"` are compared with DBpedia labels such as `"Australia"`.



```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/">
<GetRecord>
...
<record>

<header>
  <identifier>oai:nla.gov.au:nla.map-nk1476</identifier>
  <setSpec>Map</setSpec>
  ...
</header>

<metadata>
  <oai_dc:dc ...>
    <dc:title>The south eastern portion of Australia showing the
    routes of the three expeditions and the surveyed territory...
    </dc:title>
    <dc:creator>Mitchell, Thomas, Sir, 1792-1855.</dc:creator>
    <dc:coverage>1838</dc:coverage>
    <dc:date>1838</dc:date>
    <dc:description>Mitchell's map of Victoria and New South Wales
    showing towns, major rivers... </dc:description>
    <dc:subject>Mitchell, Thomas, Sir, 1792-1855 -- Travel -- Australia,
    Southeastern -- Maps.</dc:subject>
    <dc:subject>Australia, Southeastern -- Discovery and exploration
    -- Maps.</dc:subject>
    <dc:subject>Victoria -- Discovery and exploration
    -- Maps.</dc:subject>
    <dc:subject>Australia, Southeastern -- 1836 -- Maps.</dc:subject>
    <dc:subject>New South Wales -- 1836 -- Maps.</dc:subject>
    <dc:subject>Victoria -- 1836 -- Maps.</dc:subject>
    <dc:contributor>Davies, Benjamin Rees.</dc:contributor>
    <dc:contributor>Mitchell, Thomas, Sir, 1792-1855. Three expeditions
    into the interior of Eastern Australia.</dc:contributor>
    ...
  </oai_dc:dc>
</metadata>

</record>
</GetRecord>
</OAI-PMH>
```

Figure 16: Sample NLA metadata record.

4.3 Dataset 3: Austrian National Library — Dublin Core

The OAI-PMH repository of the Austrian National Library's (ONB) image archive, available at `http://oai-bdb.onb.ac.at/Script/oai2.aspx`, exposes metadata about digitised historical images, especially around the period of World War II, both in Dublin Core and a proprietary format. We used the latter for setting up an OAI2LOD Server instance and therefore had to create a specific style sheet. Knowing that many images show German politicians from the period of World War II, we have defined a linking rule that analyses the element `onbba:personDepicted`, in particular the names of these persons, against all resources in DBpedia that are categorised as `yago:GermanPeopleOfWorldWarII`.

Since persons names are represented differently in the ONB dataset (e.g., `"Speer, Albert"`) than in DBpedia (e.g., `"Albert Speer"`), general similarity metrics algorithms were not applicable. The normalised Levensthein distance of the strings `"Speer, Albert"` and `"Albert Speer"` is 0.38; lowering the string similarity threshold to such a level would cause other string pairs such as `"Speer, Albert"` and `"Hans Albers"`, which have a Levensthein distance of 0.46, to be matched as well, resulting in semantically incorrect links. For that reason, we have imple-

```
<lrule1> a oai2lod:LinkingRule;
    oai2lod:sourceType oai:Item;
    oai2lod:sourceProperty onbba:
       personDepicted;
    oai2lod:targetType yago:
       GermanPeopleOfWorldWarII;
    oai2lod:targetProperty rdfs:label;
    oai2lod:linkingProperty rdfs:seeAlso;
    oai2lod:similarityMetrics "at.ac.
       univie.mminf.oai2lod.linking.
       NameSimilarity";
    oai2lod:minSimilarity "0.79"^^xsd:
       float;
    .
```

Figure 17: Sample OAI2LOD linking rule.

mented a domain specific string metrics (`NameSimilarity`) that splits the considered strings into tokens, removes special characters such as commas and colons, and calculates the percentage of tokens common in both strings. Figure 17 shows the linking rule we applied to this dataset.

Again we have harvested 5,000 records in the ONB-specific metadata format (`tel_onbba`) using iterative `ListRecords` requests. In total, the records contained 3,517 `onbba:personDepicted` property-value pairs. With a similarity threshold of 0.79, the linking mechanism created 214 links between ONB and DBpedia resources, which were added to 226 records (16 records were annotated with more than one link). Also in this case, all links we have examined manually turned out to be correct. Again, a number of resources in the ONB dataset were annotated with names that were not correctly classified in DBpedia (e.g., `"Riefenstahl, Leni"`): this is due the fact that in the community-based Wikipedia, on which DBpedia is based, information is partially incomplete or incorrect.

4.4 Results

From the previously described experiments we can draw two conclusions: first, the technical effort of setting up an OAI2LOD Server is minimal if standard settings are applied. When metadata is exposed only using Dublin Core vocabulary, an administrator's single task is to configure the OAI-PMH endpoint URL to be wrapped and start the server. Exposing other formats than Dublin Core requires basic knowledge on XSL transformations because the OAI2LOD Server instance must be equipped with a style sheet that converts the harvested XML metadata into RDF/XML.

Second, the quality of the links produced by the OAI2LOD Server's linking mechanism depends on (*i*) the quality of the linking rules and (*ii*) the quality and semantic precision of the harvested metadata. For many institutions the application of the Simple Dublin Core format means flattening their perhaps well-structured metadata to a set of 15 elements, which results in a loss of machine-interpretable semantics. While this is necessary in the context of OAI-PMH in order to provide a minimum level of interoperability, it restricts the possibilities of linking resources across the LOD cloud. In particular, linking rules that analyse the values of properties with a broad semantics (e.g., `dc:subject`) are more likely to produce false positives than rules that operate on fields with a narrower meaning (e.g., `onbba:depictedPerson`). For OAI-PMH data providers that aim at interweaving their data sources with the Linked Data cloud, it is therefore recommended to follow the OAI-PMH specification and provide their metadata in different, more expressive formats, too. Regarding the currently applied metadata formats (see Figure 6), this is generally not yet the case.

The problem of linking metadata is a non-trivial one and the linking mechanism provided by the current OAI2LOD Server calls for further improvement. In the literature, the linking problem is referred to *name* or *entity resolution* and has been studied intensively (e.g., Benjelloun et al. (2009)). A possible enhancement of the OAI2LOD Server's linking capabilities is to enable the formulation of more complex linking rules that can be combined using various kinds of conditions so that one can take into account several source and target properties for determining if there is a semantic correspondence between two resources. This, however, is out of the scope this work.

5 Related Work

The OAI2LOD Server is a wrapper component or publishing tool that enables the integration of OAI-PMH data sources into the Linked Data cloud. As already mentioned before, there also exist publishing tools for other types of data sources: the D2R Server (Bizer and Seaborne, 2004) and Triplify Auer et al. (2009) are prominent examples for wrappers that expose data from relational databases on the Web. Pubby[21] is a linked data front-end for arbitrary SPARQL endpoints, and the RDFizers[22] provided by the SIMILE project can be applied for transforming common data formats (e.g., Email, bibliographic data, etc.) into RDF. A continuously updated list of available wrapper components is available in the Linking Open Data Wiki[23].

The problem of creating semantically valid links between resources is known as *entity resolution* or *record linkage* and has been addressed extensively in the database and data mining domain; e.g., by Benjelloun et al. (2009) and Elmagarmid et al. (2007). Much more advanced algorithms than the one presented in this paper have already been proposed; some of them focus on similarity in text (Chaudhuri et al., 2003), while others (e.g., Kalashnikov and Mehrotra (2006)) also take into account the structures of records. Often, however, algorithms tailored to a certain application domain (e.g., Raimond et al. (2008) for music metadata)

[21] Pubby: `www4.wiwiss.fu-berlin.de/pubby/`

[22] SIMILE RDFizers: `simile.mit.edu/RDFizers/`

[23] Linking Open Data Wiki: `http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/PublishingTools`

have been created and produced quite satisfactory results.

In the digital libraries and archives domain, the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) specification (Lagoze et al., 2008) is the latest standardisation effort driven by the designers of the OAI-PMH protocol. Regarding its architecture, we can notice strong similarities with the ideas of Linked Data and the OAI2LOD Server respectively. We can observe that the first two Linked Data rules are fundamental building blocks of the standard: all *things*, i.e., resource maps and the aggregated resources, are identified by dereferencable URIs. Further, all terms used for describing aggregations have a well-defined semantics, published in terms of a Web accessible vocabulary definition. It also considers the third LOD rule because resolving the URIs returns *useful*—i.e., processable and interpretable—information for both humans and machines. Finally, OAI-ORE also follows the fourth rule by providing several possibilities to link resources: first, an aggregation of resources is by definition a collection of linked resources; second, ORE defines the `ore:similarTo` property to express that an ORE aggregation is similar to a linked resource; third, it supports the concepts of nested aggregations.

6 Conclusion

In this paper we have proposed the OAI2LOD Server, a publishing tool that can wrap existing OAI-PMH data sources and expose the available items and metadata sources *on* the Web according to the Linked Data requirements. It closed the conceptual gap between the OAI-PMH and LOD worlds by assigning dereferencable URI identifiers to items and sets, their various representations (XHTML, RDF, etc.), and to the vocabulary terms used within metadata descriptions. In order to link OAI-PMH items with related resources in other data sources, the OAI2LOD Server provides a basic linking mechanism, which can be tailored to domain specific needs through the definition of linking rules. The OAI2LOD Server's SPARQL query interface provides structured and selective access to the available metadata and hence represents an alternative to the OAI-PMH batch harvesting approach.

The results of the case study presented in Section 4 have shown that, although the provided linking mechanism is still very simple and straightforward, the quality of the produced links is very high. If the source metadata are well structured and semantically well-defined, and if the applied linking algorithms are appropriate for the target domain, we can obtain highly qualitative link sets between different data sources. For data providers that aim at integrating their OAI-PMH data endpoint into the LOD cloud, we recommend from our experience to follow the OAI-PMH guidelines and expose their metadata also in other formats than Dublin Core.

Regarding the OAI-ORE developments, we can observe that the LOD principles already play an important role in the digital libraries domain. We have also seen that the conceptual gap between OAI-PMH and ORE is narrow and can easily be bridged by intermediate gateways like the OAI2LOD Server. Since the LOD approach actually subsumes a large fraction of the OAI-PMH functionalities, we believe that future releases of the OAI-PMH standard could even consider a shift to the LOD principles, which would also enable a tighter integration with the OAI-ORE protocol. Meanwhile, the OAI2LOD Server can be used for bridging the conceptual gap between these standards.

The major construction areas in the OAI2LOD Server are the linking mechanism and the persistence backend. For producing better linking results with less restricted linking rules we plan to adopt more sophisticated record linkage techniques and advanced configuration options. To increase the capacity and performance we plan to integrate the OAI2LOD Server with large scale triple-storage solutions such as OpenLink Virtuoso[24]. We will also investigate possible solutions that deal with the link sustainability problem in the context of open, uncontrolled environments such as the Web. In particular, we will analyse the effects and appropriate measures when linked resources change, move, disappear or become semantically invalid.

ACKNOWLEDGMENTS

We would like to thank our colleagues Wolfgang Jochum and Stefan Zander for their valuable comments on this paper.

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International Semantic Web Conference (ISWC 2007), Busan, Korea.*

Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., and Aumueller, D. (2009). Triplify - light-weight linked data publication from relational databases. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web.* Accepted for publication.

Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., and Widom, J. (2009). Swoosh: A Generic Approach to Entity Resolution. *VLDB J.*, 18(1):255–276.

Berners-Lee, T. (2006). *Linked Data.* World Wide Web Consortium. Available at `http://www.w3.org/DesignIssues/LinkedData.html`, retrieved 08-Aug-2008.

Berners-Lee, T., Fielding, R., and Masinter, L. (2005). *Uniform Resource Identifier (URI): Generic Syntax (RFC 3986).* Network Working Group.

[24] OpenLink Virtuoso: `http://www.openlinksw.com/virtuoso/`

Berrueta, D. and Phipps, J. (2008). *Best Practice Recipes for Publishing RDF Vocabularies (W3C Working Group Note 28 August 2008)*. World Wide Web Consortium.

Bizer, C., Cyganiak, R., and Heath, T. (2007). *How to Publish Linked Data on the Web*. Available at `http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/`.

Bizer, C. and Seaborne, A. (2004). D2RQ - Treating Non-RDF Databases as Virtual RDF Graphs. In *3rd International Semantic Web Conference (ISWC2004)*, Hiroshima, Japan. Available at: `http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/`.

Chaudhuri, S., Ganjam, K., Ganti, V., and Motwani, R. (2003). Robust and Efficient Fuzzy Match for Online Data Cleaning. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pages 313–324, New York, NY, USA. ACM.

Clark, K. G., Feigenbaum, L., and Torres, E. (2008). *SPARQL Protocol for RDF (W3C Recommendation 15 January 2008)*. World Wide Web Consortium.

DC (2006). *Dublin Core Metadata Element Set, Version 1.1*. Dublin Core Metadata Initiative. Available at: `http://dublincore.org/documents/dces/`.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Trans. on Knowl. and Data Eng.*, 19(1):1–16.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and Berners-Lee, T. (1999). *Hypertext Transfer Protocol – HTTP/1.1 (RFC 2616)*. Network Working Group.

Haslhofer, B. and Schandl, B. (2008). The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data. In *International Workshop on Linked Data on the Web (LDOW2008), co-located with WWW 2008*, Beijing, China.

ISO TC 154 (2004). *Data Elements and Interchange Formats — Information Exchange — Representation of Dates and Times — ISO 8601:2004*. International Standardizaton Organization (ISO). Available at: `http://www.iso.org/iso/catalogue_detail?csnumber=40874`.

Jacobs, I. and Walsh, N. (2004). Architecture of the World Wide Web, Volume One. Available at: `http://www.w3.org/TR/webarch/`.

Jentzsch, A. (2009). *The Linking Open Data Dataset Cloud (as of March 2009)*. Available at `http://www4.wiwiss.fu-berlin.de/bizer/pub/lod-datasets_2009-02-27.png`.

Kalashnikov, D. V. and Mehrotra, S. (2006). Domain-independent Data Cleaning via Analysis of Entity-Relationship Graph. *ACM Trans. Database Syst.*, 31(2):716–767.

Klyne, G. and Carroll, J. J. (2004). *Resource Description Framework (RDF): Concepts and Abstract Syntax (W3C Recommendation 10 February 2004)*. World Wide Web Consortium.

Kobilarov, G. and Dickinson, I. (2008). Humboldt: Exploring Linked Data. In *International Workshop on Linked Data on the Web (LDOW2008), co-located with WWW 2008*, Beijing, China.

Lagoze, C. and van de Sompel, H. (2002). The Open Archives Initiative Protocol for Metadata Harvesting — Version 2.0. Available at: `http://www.openarchives.org/OAI/openarchivesprotocol.html`.

Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M. L., Sanderson, R., and Warner, S. (2008). *Open Archives Initative Object Reuse and Exchange (OAI-ORE)*. Available at: `http://www.openarchives.org/ore/`.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10.

Malmsten, M. (2008). Making a Library Catalogue Part of the Semantic Web. In Greenberg, J. and Klas, W., editors, *Proceedings of the International Conference on Dublin Core and Metadata Applications*.

Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., and Tummarello, G. (2008). Sindice.com: A Document-oriented Lookup Index for Open Linked Data. *Internal Journal of Metadata, Semantics and Ontologies*, 3(1):37–52.

Prud'hommeaux, E. and Seaborne, A. (2008). *SPARQL Query Language for RDF (W3C Recommendation 15 January 2008)*. World Wide Web Consortium.

Raimond, Y., Sutton, C., and Sandler, M. (2008). Automatic Interlinking of Music Datasets on the Semantic Web. In *International Workshop on Linked Data on the Web (LDOW2008), co-located with WWW 2008*, Beijing, China.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). YAGO: A Core of Semantic Knowledge. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pages 697–706, New York, NY, USA. ACM.

Summers, E., Isaac, A., Redding, C., and Krech, D. (2008). LCSH, SKOS and Linked Data. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*.