

OAI-P2P: A Peer-to-Peer Network for Open Archives

Benjamin Ahlborn

Technical Information Library Hannover
benjamin.ahlborn@tib.uni-hannover.de

Wolfgang Nejd

Learning Lab Lower Saxony
nejdl@learninglab.de

Wolf Siberski

Learning Lab Lower Saxony
siberski@learninglab.de

Abstract

OAI is designed with a low-barrier technology approach, thus allowing institutions to provide content metadata with little effort. On the other hand, search capabilities are very limited on OAI data providers, and have to be provided by separate service providers. We propose that data providers form a peer-to-peer network which supports distributed search over all connected metadata repositories. Such an approach is already implemented for learning content metadata (project 'Edutella'). We describe how this technology could be reused in the OAI context. This would allow OAI repositories to provide distributed search capabilities and effortless integration of new archives within a peer-to-peer network with little additional implementation effort.

1. Introduction

In the last decade, information technology has brought publishing power to the scientist's desktop. Word processing and desktop publishing have enabled scientists to produce high quality output on their desktop computers, database tools and cheap mass storage have made it possible to build local digital libraries. Finally, networking infrastructure is providing the technical means to share digital library resources across the internet. However, this last step has yet to be completed.

Until recently, providing interoperability for digital and print libraries has been limited to the big players; university library systems, scientific publishers and library network cooperatives have the size and the resources to push proprietary protocols or implement large footprint standards like Z39.50. Smaller institutions or individual researchers do not command these resources. Search engine technology has made great headway, but using search engines to discover digital library resources is cumbersome and runs the risk of drowning in the flood of non-relevant hits while omitting the so called "hidden web", i.e. documents which can only be accessed via dynamic links generated from databases.

1.1. Open Archives Initiative: interface for metadata exchange and harvesting

This is where recent initiatives like the Open Archives Initiative (OAI) step in. In order to achieve technical interoperability among distributed archives OAI has created a protocol (Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH) based on the standard technologies HTTP and XML as well as the Dublin Core metadata scheme [1]. In addition to bibliographic schemes like RFC1807 and MARC which excel in describing documents in the "traditional" print paradigm, OAI presently supports the multipurpose resource description standard Dublin Core. Dublin Core is both simple to use and versatile, although it is too general to supply fine-grained information. However, OAI-PMH itself is just providing a XML-wrapper for metadata and can be adapted to both simple and complex metadata sets.

One of the main aims of OAI was to keep the protocol simple and easy to implement. The positive feedback and rapid adoption of the OAI-PMH by the scientific communities and information professionals have proved this concept right: the number of OAI-enabled repositories is increasing, incorporating smaller repositories and institutions as well as areas of science which have not been represented in the earlier attempts.

The OAI-PMH is a protocol limited to incremental metadata transfer, providing a technical and organizational framework / environment for metadata harvesting. To keep the instruction set simple, OAI-PMH calls for a separation between data and service providers. Data providers establish an OAI-PMH-based interface to local digital resources, while service providers (like ARC [2] and SCIRIUS [3]) provide facilities for searching across multiple archives plus value-added features like ranking and unified access to other sources.

This separation exposes the simplicity of the protocol as the source of its' strength (low barrier to adoption) and its' weakness: OAI-PMH is designed as simple as possible for data providers at the expense of service providers; creating and maintaining an OAI-PMH service provider requires much more resources than setting up a data provider. On the other hand, OAI-PMH offers no front-end services: data providers offer an interface for

metadata harvesting to outsiders but do not have any immediate advantages (like a query service for outside repositories) from their efforts, unless a service provider provides an interface to their data.

1.2. OAI: enhanced concepts and solutions

Inside and outside the OAI community there have been several promising approaches to solve this dilemma. Most notable is the Kepler project bringing OAI-data provider functionality to the publishing individual. Kepler provides OAI “out of the box”-tools and a networking framework which scales up to 10.000 small repositories (e.g. single persons, small research institutes) [4]. Main features are

- a JAVA-‘archivlet’ which installs on the client’s computer to handle user data, registration with central server, metadata entry form to create OAI-compliant metadata and resource management
- a LDAP-based network environment including automated registration service, keeping track of connected clients, harvesting of clients’ metadata
- a query/discovery service (using OAI-service ARC), which provides caching of offline clients’ resources and also provides services for general users outside the Kepler framework.

Kepler succeeds in bringing services to the data providers while preserving technical simplicity and usability but still relies on a central service provider. The data providers have been scaled down to the individual and attached to one single service provider. Apart from the concept of sets in OAI-PMH, Kepler does not support community building.

A similar concept tailored for a clear-cut audience but not based on OAI-PMH is the *JOINed* Digital Library (JINI Object Information Network) [5].

1.3. Edutella: a peer-to-peer metadata infrastructure

Recently peer-to-peer systems have evolved as an alternative concept for digital resource sharing. In contrary to traditional client/server systems in a peer-to-peer network each participating peer (network node) can be consumer and provider of data and services [6]. Therefore it becomes much easier for a participating party to become provider of information.

Edutella is a peer-to-peer infrastructure for storing, querying and exchanging metadata [7]. It is built on the open source project JXTA, a framework which provides basic peer-to-peer network features [8]. Edutella connects highly heterogeneous peers (heterogeneous in their uptime, performance, storage size, functionality, number of users etc.). To achieve the desired interoperability, it is crucial to adhere to standards [9]. Therefore Edutella is based on metadata standards defined by the SemanticWeb

initiative of the WWW Consortium [10], namely RDF and RDFS. Each Edutella peer can make its metadata information available as a set of RDF statements, suitable for describing distributed resources.

Peers publish what they offer by announcing which kind of services they provide. In the OAI context, the most important services are query and replication service.

The Edutella query service is the most basic service within the Edutella network. Just like with the OAI “ListMetadataFormats” request, peers register the queries they may be able to answer through the query service (i.e., by specifying supported metadata schemas (e.g., “this peer provides metadata according to the DCMI standards”). Queries are sent through the Edutella network to the subset of peers who can potentially deliver results. The resulting RDF statements are sent back to the requesting peer.

Each peer is free to use its own language for query processing. However, there must be a common language to facilitate query distribution. Therefore Edutella defines a family of query exchange languages (QEL) based on a common datamodel, starting with simple conjunctive queries (which allow a query-by-example style of request) up to query languages equivalent to query languages of state-of-the-art relational databases.

The same applies to the query front-end. It was straightforward to implement a form based query front-end which translates the input into QEL before sending the request to the peer network. There is also a graphical query editor available called Conzilla [11].

Each QEL query is based on explicitly referenced metadata schemas (e.g. DC, LOM) and is independent of a specific schema.

To make life as easy as possible for peer implementers, Edutella employs a plug-in architecture where a new peer only needs to provide a translator between QEL and its own query language and a query processor for its data store. Everything else is handled by the Edutella

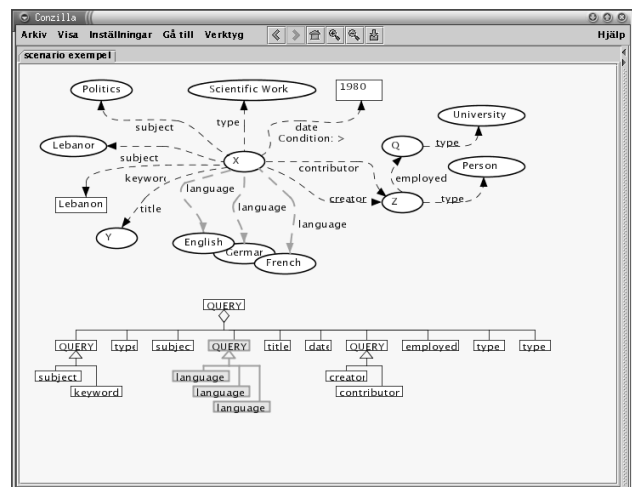


Figure 1. Conzilla as query editor

framework.

The replication service (currently under development) is complementing local storage by replicating data in additional peers to achieve higher reliability and workload balancing while maintaining data integrity and consistency. It also allows higher availability of metadata of smaller peers when they replicate their data to a peer which is always online.

Another part of the Edutella project is the implementation of mapping services which will allow translating between different schemas (e.g. from MARC to DC)

2. OAI-P2P: Extending OAI with peer-to-peer concepts

In essence, the OAI-PMH defines client-server-relationships, with the data provider and the service provider responding to client requests. However, digital libraries in most cases act both as a client and as a server, at the same time trying to obtain outside material for inside users and offering inside resources to outside users. Similar initial situations have spawned the emergence of services like Napster and Gnutella, which offer resource sharing by means of peer-to-peer structures. This paper describes an organizational and technical framework which merges the OAI-PMH concept with a true peer-to-peer approach (OAI-P2P). It thus takes the OAI-PMH one step further by extending query services to data providers and by avoiding the dependencies of centralized server-based systems.

2.1. Motivation

Figure 2 shows a typical OAI topology. Different data providers are harvested by different service providers. All queries are handled by providers at the service provider layer.

When a user wants to query all data providers, he has to send a query to multiple service providers. The results will overlap, and the client will have to handle duplicates.

Note also that this architecture makes it difficult for a new data provider to get accessible. As long as no service provider is willing to harvest its metadata, end user won't 'see' them.

Another issue occurs when service providers are terminated or reorganized. The most prominent example is Networked Computer Science Technical Reference Library (NCSTRL): the service suffered from limited availability for the best part of 2000 and 2001 (according to [12] due to funding problems). In such a case, the data providers attached to this service provider may find that their archive is no longer harvested, and they lose

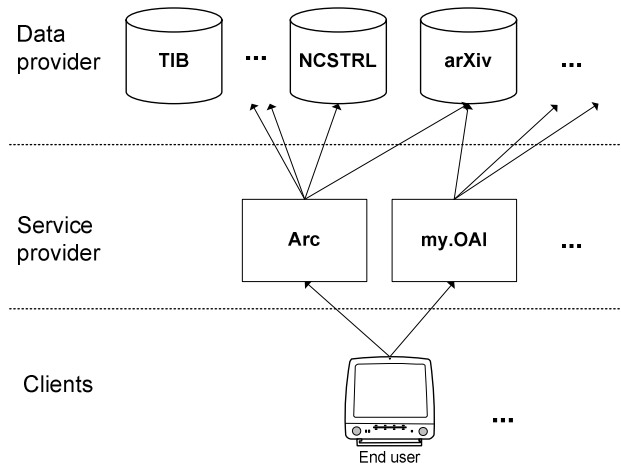


Figure 2. OAI topology

access to other repositories formerly made accessible by the discontinued service provider. The whole infrastructure has to be re-established with a new service provider.

In a P2P-system, there is no separation between service provider and data provider (each peer maintains separate subsystems for data storage and query handling). Each query is routed to appropriate peers by the network; there is no administration necessary to introduce new peers. Of course such a network still benefits from additional service providers which replicate metadata, thereby enhancing the reliability and performance of the net. However, although performance may suffer if such a peer is discontinued, overall communication and services will stay alive even if a single node dies.

Community building is not a technical but a social process. Individual digital libraries may want to decide which other repositories they get to share their data with

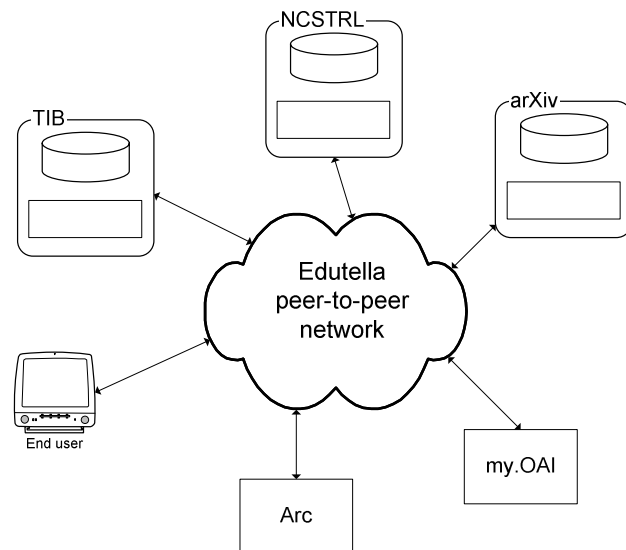


Figure 3. OAI-P2P topology

and which repositories they want to access. In the present OAI framework, choosing the scope of the community is the prerogative of the service provider, who may arbitrarily decide which data providers to include. With the P2P approach peers can devise community specific access policies using the peer group concept. Under current client-server conditions, a data provider has to build a new service provider from scratch if the targeted community is not featured by existing service providers. Under a peer-to-peer architecture, data providers take advantage of already existing infrastructure to create community specific services by introducing a new peer group.

The OAI-PMH is “pull”-based, i.e. it relies on the service provider to perform regular metadata harvests, thus leaving the client in a state of possible metadata inconsistency. OAI-P2P allows data providing peers to “push” their data, thereby making sure that all interested peers receive timely and concurrent updates, keeping the peer group synchronized.

In general, our motivation behind implementing a P2P-based solution extending OAI-PMH (OAI-P2P) is to enhance resource sharing as described above while maintaining the OAI-PMH’s strong points, i.e.

- Support for heterogeneous backend repositories: peers decide how to set up and organize repository servers.
- Interoperability: OAI-P2P defines interfaces for communication with other repositories.
- Adaptability (both to different platforms and community-specific metadata sets and information needs).
- Low barrier to implementation.

2.2. Query capabilities

Merging data provider and service provider functionality means that an OAI-based peer-to-peer digital library network has to address issues outside the OAI-PMH scope. OAI-PMH does not state how data providers should set up source metadata. Although very small archives can use the file system to store XML-metadata, most institutional data providers use a dedicated relational database from which OAI output is created. On the other hand, current OAI service providers replicate the metadata they have harvested in relational databases to provide for clients’ queries. In order to give service provider functionality to each data provider in a peer-to-peer context, repositories must be able to pose, process and accept advanced queries, an ability which they use anyway to build an OAI-compliant infrastructure in the first place.

By providing nothing but a container for metadata the OAI-PMH evades the issue of metadata content and structure, leaving it as a problem for service providers to address. This means OAI-P2P-implementations have to

support queries to extract metadata information. At present, this often means searching flat tables containing bibliographic metadata. However, metadata are bound to become more complex, incorporating links and references to additional data, e.g.

- Terms and conditions of full-text use, local licensing agreements. Terms-and-conditions vocabularies are for the most part work-in-progress but we expect machine readable schemes for these metadata in the near future.
- Authority file headers (e.g. community-specific classification tags)
- Document hierarchy: information about supplementary material (field data, visualizations), links to related documents etc.
- Peer review information (annotation, version control)

Using Edutella’s support for a range of query exchange languages (QEL), OAI-P2P is able to adapt to heterogeneous peers as well as changing demand on metadata scope and query complexity.

2.3. Scenario

What kind of services could be built on top of an OAI-P2P implementation? Let us assume a scenario where a research institute has decided to share digital resources with the scientific community. In a first step, an OAI-compliant metadata infrastructure has been set up. The enhanced Edutella-software is downloaded and installs on top of the OAI-framework, transparently providing instant basic services like request handling, peer registration, provider and consumer service search. The first registration with the peer-to-peer network kicks off a message to all registered peers containing the OAI-“identify”-statement, declaring their intended query spaces and what sort of queries they wish to respond to. Depending on community-specific practices and vocabularies, this statement may contain keyword or other metadata which will in turn generate a response of several “Identify”-statements to the newcomer repository. Thus being notified that a new digital library is available, other peers may add the new resource to their community list. In the same way, a new peer may also actively query available nodes to detect fellow peers. Another way to discover peers is using resource queries. A community-specific query is directed to all available archives. Those providers who are able to return results are added to the list of peers. If not explicitly stated, subsequent queries are always directed to this list of peers. If a query transcends the community’s scope, it may be extended to all available peers or to other specific peer groups. This list can of course be edited manually.

Resource discovery is of course the core service of OAI-P2P. Each query to a provider peer triggers an OAI-compliant response which is returned to the consumer

peer. Depending on the OAI-metadata infrastructure, all or a part of the responses may be cached or discarded after the session. In most cases data will be added to the local peer's database or file-system. As a default, queries are only executed on metadata for which the peer is directly responsible; in case of community members with unreliable uptimes queries may be extended to cached data, with the OAI identifier pointing to the original source. Inside OAI-P2P communities or hubs, new resources may be broadcasted to all peers, thus "pushing" instant updates to peer databases or caches. After initialising a new peer by harvesting the metadata regarded useful the process of updating inside the chosen peer community is automatic. As every peer is not only data consumer but also data provider, the infrastructure to process and store OAI-compliant data is readily available.

After its inclusion in a subject based community the newcomer's users start putting requests to the other repositories. At present, responses are flat hierarchical metadata. As the OAI-PMH is providing only the container for metadata schemes, OAI responses may also contain links to other resources, e.g.

- technical papers reporting progress in engineering may contain a pointer to CAD objects which can be downloaded.
- documents from other fields may contain links to supplementary material like large volume measurement data
- intellectual property language.
- courseware

P2P-OAI adopters take advantage of Edutella's built-in SQL-like query interface which adapts to complex data structures. Depending on the type of resource, further services like peer review or resource annotation can be used [13].

3. Design considerations

3.1. System architecture

There are two variants to enable an OAI data provider to become an OAI-P2P peer.

The first variant is to wrap the provider with a peer which replicates the data to an RDF repository. For small peers (less than 1000 documents) an RDF file would suffice as repository. This solution is appropriate if either the amount of data is small or it is difficult to access the data directly. Such a peer can make content available from several data providers and is very similar to a service provider in the classical sense of OAI. This peer type is therefore also suited to integrate arbitrary OAI data providers into OAI-P2P. Such an OAI-P2P wrapper (which we call data wrapper, because it only uses the data providing capabilities from the underlying repository) is

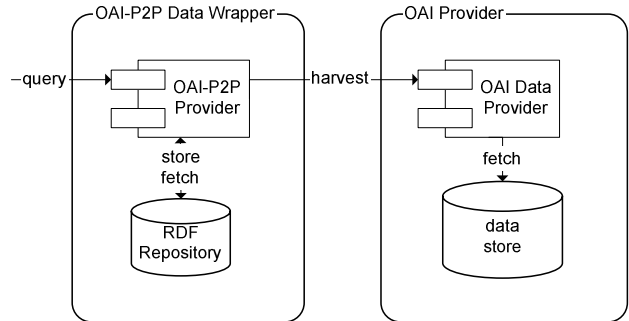


Figure 4. wrapper for OAI data provider

implemented once, and only has to be configured to provide the query service for specific data providers.

The second variant is to answer queries directly from the data provider's database. In this case, the new peer interface needs to transform the QEL query to a query understandable by the underlying data store. We call this an OAI-P2P query wrapper, because it wraps the underlying repository including its query capabilities. This solution doesn't need to replicate data and therefore ensures that the query response is always up-to-date. It may also improve performance. On the other hand such a peer has to be developed for each type of data store.

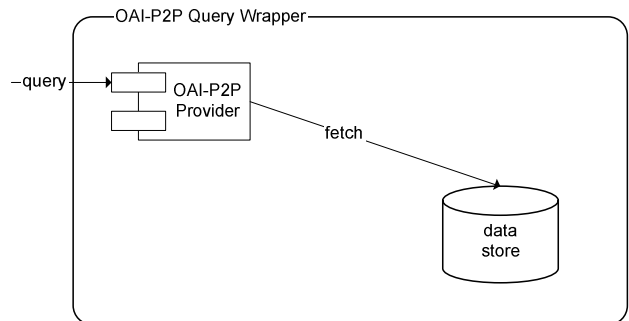


Figure 5. peer accessing data directly

3.2. Message format

As all data within the Edutella network is transported in RDF format, we need to define an RDF-Binding for OAI. This has already been done for Dublin Core [14]. We only need to add OAI specific information.

For example, a response to a query equivalent to the OAI Get request would now look like the following (namespace declarations omitted):

```
<oai:Result>
  <oai:responseDate>
    2001-06-10T14:09:57-07:00
  </oai:responseDate>
  <oai:hasRecord rdf:resource=
    "http://arXiv.org/.../9901001"/>
</oai:Result>
```

```

<oai:Record rdf:about=
  "http://arXiv.org/.../9901001">
  <dc:title>Quantum slow motion</dc:title>
  <dc:creator>Hug, M.</dc:creator>
  <dc:creator>Milburn, G. J.</dc:creator>
  <dc:description>We simulate the center of
    mass motion of cold atoms in a
    standing, amplitude modulated, laser
    field as an example of a system that
    has a classical mixed phase-space.
  </dc:description>
  <dc:date>1999-01-01</dc:date>
  <dc:type>e-print</dc:type>
</oai:Record>

```

As proof of concept we will develop two OAI-P2P prototypes in the next months, each exploiting one of the design variants. This will extend the current Edutella peers which have been geared towards learning material repositories using different RDF and XML repository query languages (RDQL, O-Telos, SQL, XPath, etc.) to peers which (as mediators) provide digital library content.

4. Conclusion

We have described how introducing a peer-to-peer approach to OAI remedies some of the current deficiencies and opens new opportunities to open archive collaboration. Although the effort in terms of technology use would be larger than the existing OAI-PMH, a lot of the necessary technology can be provided by using an already existing framework, Edutella, and the extended OAI-P2P network can easily include existing OAI-PMH services using combined OAI-PMH / OAI-P2P service providers. Therefore the technology barrier to OAI-P2P stays reasonably low. We think that giving OAI data providers the benefits of effortless integration of new archives in their peer community as well as the enhanced query capabilities are easily worth the additional overhead.

5. References

- [1] C. Lagoze, and H. Van de Sompel, "The open archives initiative: building a low-barrier interoperability framework", *ACM/IEEE Joint Conference on Digital Libraries*, 2001.
- [2] X. Liu, K. Maly, M. Zubair, and M. L. Nelson, "Arc: an OAI service provider for cross-archive searching", *ACM/IEEE Joint Conference on Digital Libraries*, 2001.
- [3] See http://www.scirus.com/html/scirus_service_provider.htm
- [4] X. Liu, K. Maly, and M. Zubair, "Kepler - An OAI Data/Service Provider for the Individual", *D-Lib Magazine*, April 2001. Available at: <http://www.dlib.org/dlib/april01/maly04maly.html>
- [5] See <http://invision.gsfc.nasa.gov/join/JDL/html/index.htm>
- [6] Andy Oram (Ed.), "Peer-to-Peer: Harnessing the Power of Disruptive Technologies", O'Reilly, 2001
- [7] W. Nejdl, B. Wolf, C. Qu_, S. Decker_, M. Sintek, A. Naeve, M.Nilsson, M. Palmér_ and T. Risch, "Edutella: A P2P Networking Infrastructure Based on RDF". Accepted for *Eleventh International World Wide Web Conference*, 2002. Available at: <http://edutella.jxta.org/reports/edutella-whitepaper.pdf>.
- [8] L. Gong, "Project JXTA: A Technology Overview", Sun Microsystems, 2001. Available at <http://www.jxta.org/project/www/docs/TechOverview.pdf>.
- [9] R. Dornfest, and D. Brickley, "The Power of Metadata", in *Peer-to-Peer: Harnessing the Power of Disruptive Technologies* [6]. Available at <http://www.openp2p.com/pub/a/p2p/2001/01/18/metadata.html>
- [10] T. Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web", *Scientific American*, May 2001. Available at <http://www.sciam.com/2001/0501issue/0501berners-lee.html>.
- [11] A. Naeve, "The Concept Browser - a new form of Knowledge Management Tool", *Proceedings of the 2nd European Web-based Learning Environments Conference (WBLE 2001)*, 2001.
- [12] T. Krichel, S. M. Warner, "Academic self-documentation: which way forward for computing, library and information science?", *ICADL 2001*. Available at: <http://openlib.org/home/krichel/mitaka.a4.pdf>.
- [13] W. Nejdl, B. Wolf, S. Staab, J. Tane, "EDUTELLA: Searching and Annotating Resources within an RDF-based P2P Network", Accepted for *International Workshop on the Semantic Web, Eleventh International World Wide Web Conference*, 2002. Available at http://edutella.jxta.org/reports/edutella_p2p.pdf.
- [14] D. Beckett, E. Miller, and D. Brickley, "Expressing Simple Dublin Core in RDF/XML", *DCMI Proposed Recommendation*, 2001. Available at <http://dublincore.org/documents/2001/11/28/dcmes-xml/>.