Преглед НЦД 21 (2012), 19–24

Bojan Marinković, Zoran Ognjanović
Mathematical Institute
of the Serbian Academy of Sciences and Arts
Tamara Butigan Vučaj
National Library of Serbia

A DISTRIBUTED IMPLEMENTATION OF A CATALOG OF DIGITIZED CULTURAL COLLECTIONS

Abstract: We present a distributed catalog of digitized cultural collections. The implementation is based on the recommendation for the national standard for describing collections and the distributed hash table (DHT) protocol Chord. DHT-based networks have emerged recently as a flexible decentralized solution to handle large amount of data without the use of high-end servers. Data storage and data retrieval from different kind of content providers (libraries, archives, museums, universities, research centers, etc.) can be handled uniformly inside a distributed catalog, such that every provider is responsible for its own part of information. This approach can be extended to allow interconnection of different DHT-based networks using an overlay network protocol called Synapse.

Keywords: digital collections, distributed protocol, standardization

1. Introduction

The document "Recommendations for coordination of digitization of cultural heritage in South-Eastern Europe (SEE)" [4] argues that the current digitization practice in SEE still does not match the priorities communicated on the EU-level and says that "It is recognized that knowledge of the cultural and scientific heritage is essential for taking decisions concerning its digitization and for interpreting the digitized resources. For this reason, inventorying and cataloging should precede or accompany the digitization of cultural and scientific assets." From the other side, SEE digital collections are still not included enough in widespread digital libraries, like Europeana, World digital library, Manuscriptorium, so there is a need for connecting these collections in the SEE network.

The initiative of the National Center for Digitization (NCD) in Serbia recognized in the very beginning of its existing that the metadata issue is among the most sophisticated ones in the complete digitization process. That's why the Center created a metadata working group to develop the metadata schemas for describing digitized heritage, pretending to be the national standards. This work was encouraged by the lack of the wide spread metadata standard for describing digitized heritage in Serbia. *The recommendation of the metadata schema for the mobile heritage digitized objects* was generated in 2007, covering archives, library, museum objects [5,7]. Many of the existing metadata standards (Dublin Core, EAD, MARC) were considered during creation of this schema. The definition of Application Profile as an "assemblage of metadata elements selected from one or more metadata and combined in a compound schema" according to Dublin Core Initiative perfectly suits the new defined schema. The European Library Application Profile for objects was a sort of inspiration for NCD metadata model as well as for some basic FRBR concepts.

Digital objects are usually organized in digital collections. A considerable work has gone into building aggregations of digital collections, which was the case with the Digital National Library of Serbia too [6]. The logical extension of the NCD metadata working group activity was towards collection-level descriptions. While some of the well-known aggregations like Europeana [11] or OAIster [12] do not use collection-level metadata, NCD supports the idea of presenting cultural and scientific heritage aggregated in digital collections. In 2009, the efforts of NCD metadata working group resulted in a new *Proposal of metadata schema for digital collection description*. For this schema, The European Library Application Profile for collections [10] was the starting point, as well as the MICHAEL collection description [8]. This schema contains descriptive (17 elements) and administrative (5 elements) metadata. It allows one to create a catalog containing descriptions of physical and/or virtual collections of the already digitized objects.

The metadata area is not resistant to Web 2.0 influences; on the contrary, it's been radically changing. Changes like social tagging, indexing and annotation bring more democracy in this traditionally conservative domain, including broader community in the kingdom of memory institutions. That's why it is important to have the technical environment supporting the aggregation of all available metadata and providing an open space for exchanging metadata. In this context, the NCD recommendations for metadata schemas are just the starting point to describe the heritage, leaving to the users to adjust it according to their own needs.

In this paper we present a distributed catalog of descriptions of digital collections. The implementation is based on the Chord protocol [3]. In Chord data storage and data retrieval from different kind of content providers (libraries, archives, museums, universities, research centers, etc.) can be handled uniformly inside one distributed catalog, such that every provider is responsible for its own part of information. We note that a possible extension of this approach can be naturally realized in the form of a regional network of the professional/national distributed catalogs. This can be achieved using an overlay network protocol called Synapse [2, 9].

2. Description of the standard

Our schema contains two groups of metadata elements:

- descriptive (17 elements) and
- administrative (5 elements),

listed below (mandatory fields are marked with asterisks, while ... denotes that the corresponding fields can be repeated):

- descriptive:
  - title*
    - original title*
    - title version ...
      - title version language*
      - title version*
  - creator ...
    - identifier
    - role ...
  - contributor ...
    - identifier
    - role ...

    - owner
    - subject ...
    - classification ...
        - classification scheme*
        - classification identifier*
    - description ...
        - source language
        - source
    - period of existence ...
        - date of creation
        - date of dismisson
        - comment ...
            - source language
            - source
    - type*
    - nature of collection
    - identifier*
    - collection's object ...
    - history ...
        - source language
        - source
    - related collection ...
        - type of relation*
        - identifier*
    - source object id
    - bibliography ...
    - note ...
        - source language
        - source
- administrative:
    - rights
    - access rights
    - record creation date
    - record creator
    - record owner

3. Chord protocol

The distributed hash table (DHT) based networks have emerged recently as a flexible decentralized solution to handle large amount of data without use of high-end servers. One of the most popular DHT-protocol is Chord in which a number of nodes is running the protocol form a ring-shaped network. The main operation supported by Chord is:

- mapping the given key onto a node using consistent hashing.

The consistent hashing provides load-balancing, i.e., every node receives roughly the same number of keys, and little movement of keys is required when nodes join and leave the network. Chord networks are overlay systems. Thus, each node in a network (with N nodes) needs "routing" information about only a few other nodes ($O(\log N)$), and resolves all lookups via $O(\log N)$ messages to other nodes. When the network is not stable, i.e., the corresponding

"routing" information is out of date since nodes join and leave arbitrarily, the performance degrades. However, Chord's stabilization algorithm (with minor modifications) maintains good lookup performance despite continuous failure and joining of nodes.

Identifiers are assigned to nodes and keys by the consistent hash function. The identifier for a node or a key, hash(node) or hash(key), is produced by hashing the node's IP address (the value of the key). The length of identifiers (let's say of m bits) must guarantee that the probability that two objects of the same type are assigned same identifiers is negligible. Identifiers are ordered in an identifier circle modulo 2^m. Then, the key k is assigned to the node such that hash(node) = hash (key). If such a node does not exist, the key is assigned to the first node in the circle whose identifier is greater than hash(key).

Every node possesses information on its current successor and predecessor nodes in the identifier circle. To accelerate the lookup procedure, a node also maintains routing information in the form of the so-called Finger Table with up to m entries. The i^{th} entry in the table at the node n contains the identifier of the first node s that succeeds n by at least 2^{i-1} in the identifier circle, i.e., s = successor($n + 2^{i-1}$) , where $1 \leq i \leq m$ (and all arithmetic is preformed modulo 2^m). The stabilization procedure implemented by Chord must guarantee that each node's successor pointer and finger table are up to date. The procedure runs periodically in the background at each node. To increase robustness, each Chord node can create a successor list of size r, containing the node's first r successors.

Beside the mapping of keys onto the set of nodes, the only other operations realized by Chord are:

- adding/removing of a node to/from a network.

When a node n joins an existing network, certain keys previously assigned to n's successor now become assigned to n. When node n leaves the network regularly, it notifies its predecessor and successor and reassigns all of its keys to the successor.

4. Application Description

Our implementation of the proposed distributed catalog is based on the open-chord v. 1.0.5 implementation, developed by Distributed and Mobile Systems Group Lehrstuhl für Praktische Informatik Universität Bamberg, which is a Java implementation of the Chord protocol. We developed a graphical user interface (GUI) in Java distinguishing two types of users: the first ones – providers – that store information about their collections and the others – visitors – that can search a catalog. Note that only providers are nodes in our Chord network, while visitors' access is realized through providers' portals.

The catalog contains records which follow the mentioned recommendation of the meta-data format. In this implementation the following fields are searchable: original title**,** creator's identifier, owner, classification scheme and identifier**.** They are stored as in Table 1. For every searchable field the pair:

- field identifier and the corresponding value,
- hash value of the entire metadata record

is stored (rows 1 to 4 in Table 1). The catalog also contains pairs of the form:

- hash value of the entire metadata record,
- metadata record

for every collection. Basic editing features, like saving and loading records to/ from an XML file, copying/pasting and printing the XML raw data, are provided.

The search mechanism has two phases. During the first phase, having a key value (the first kind of entries) we attempt to find the hashed value of the corresponding metadata record

and then, during the second phase, to find the entire metadata record (the second kind of entries).

No.	Key	Value
1	original title#Title	hash(◊)
2	creator#Id	hash(◊)
3	owner#Owner	hash(◊)
4	classification #Sheme#Id	hash(◊)
5	hash(◊)	◊

Table 1: Different data structures for a collection stored in the catalog
(◊ represents full metadata record of a collection)

Interested readers can experience the implementation at the address:
http://arkanoid.mi.sanu.ac.rs/collections_catalog [31.12.2011].

5. Conclusions

We have presented a distributed catalog of digital collections. We have described a concrete realization of this idea which relies on the distributed hash table protocol Chord and the proposed metadata schema for describing collections. (Digitized) collections, by their nature, are highly distributed resources. Researches and members of a wider community will benefit from connecting different kinds of providers into one system, because information from a number of sources will appear at one place. An interesting possibility, which is a challenge for further work, is related to an application of the Synapse protocol in developing of a regional network of the professional/national distributed catalogs.

Acknowledgements. The work presented here was partially supported by the Serbian Ministry of Education and Science (project III44006).

References

[1] *A Recommendation for the National Standard for Describing Collections in Serbia* NCD, UNESCO Committee for digitization, http://www.ncd.org.rs/ncd_sr/standards/opis_kolekcija.html (31.12.2011)
[2] L. Liquori, C. Tedeschi, L. Vanni, F. Bongiovanni, V. Ciancaglini and B. Marinković: *Synapse: A Scalable Protocol for Interconnecting Heterogeneous Overlay Networks*, Lecture Notes in Computer Science, vol. 6091 (p. 410), p. 67–82, Edited by: Crovella, Mark; Feeney, Laura Marie; Raghavan, S.V., Springer Berlin / Heidelberg, ISBN: 978-3-642-12962-9, 2010
[3] I. Stoica, R. Morris, D. Karger, M. Kaashoek, H. Balakrishnan, *Chord: A Scalable Peer-to-Peer Lookup service for Internet Applications*. In: ACM SIGCOMM, pp. 149–160, 2001.
[4] Recommendations for coordination of digitization of cultural heritage in South-Eastern Europe, Conclusions of the Regional Meeting on Digitization of Cultural Heritage, Ohrid, Macedonia, 17–20 March 2005, Review of the National Center for Digitization, 2–7, 2005. (http://elib.mi.sanu.ac.rs/-files/journals/ncd/7/ncd07002.pdf (31.12.2011)
[5] A Recommendation of the metadata schema for the mobile heritage digitized objects, UNESCO Committee for digitization, http://www.ncd.matf.bg.ac.rs/?page=news&lang=sr&file=predlog-StandardaMetadata.htm (31.12.2011)
[6] Digital National Library of Serbia, www.digital.nb.rs (31.12.2011)
[7] Z. Ognjanović, T. Butigan, B. Marinković, NCD Recommendation for the National Standard for Describing Digitised Heritage in Serbia, in: Metadata and Semantics, eds: M.-A. Sicilia, M. D. Lytras, Springer, 45–54, 2009.
[8] MICHAEL collection description, http://www.mla.gov.uk/what/publications/~/media/Files/pdf/2007/-michael_manual_v2.ashx (31.12.2011)

[9] B. Marinković, L. Liquori, V. Ciancaglini and Z. Ognjanović: A Distributed Catalog for Digitized Cultural Heritage, ICT Innovations 2010, CCIS 83 (p. 378), Edited by: Gusev, Marjan; Mitrevski Pece, Springer-Verlag Berlin Heidelberg, 176–186, 2011.
[10] TEL Application Profile for collection descriptions version 1.3, http://www.theeuropeanlibrary.org/-portal/organisation/cooperation_old/archive/telproject_archive/tel_ap_cld_v1.3.html (31.12.2011)
[11] Europeana http://www.europeana.eu (31.12.2011)
[12] OAIster http://www.oclc.org/oaister/ (31.12.2011)

bojanm@mi.sanu.ac.rs
zorano@mi.sanu.ac.rs
tamara@nb.rs