Persistent identifiers for the national bibliography context

NUSL seminar, 22. 10. 2009, BRNO

Ladislav Cubr, Jan Hutar, Marek Melichar
National Library of the Czech Republic

Introduction

The aim of this paper is to provide an overview of the strategy in the area of the persistent identification (primarily for the national bibliography purposes) outlined at the National library in the Czech Republic (thereinafter NLCR). Besides, it is suggested that all institutions and groups interested in implementation of the national URN:NBN:CZ system should run a project, which would specify needs of participating institutions, come up with a general agreement at the conceptual level, solve the administrative issues, formulate functional requirements for the system components, and run the program's pilot. After a thorough analysis of the URN:NBN implementations in other countries, we have realized that such project will be inevitable if we really want to move to persistent, unequivocal identification system, which will also work as a resolution service and fit well into existing systems of digital libraries, catalogs and long-term preservation archives in the country.

Historia est vitae magistra

Conceptualizing persistent identifiers for the digital world has rather a long history in Czech Republic. Internet archeology presents us with a clear evidence that the Czech library community was exposed to original ideas of the URN:NBN identification system's founding father as early as 2001. Juha Hakala made a very clear statement at the CASLIN seminar in Beroun. The topic hasn't lost its relevance for the current Czech situation. Hakala [1] summarized all the basic issues the digital world brought into the identification systems set up in a traditional context. However, there was one crucial point in the above-mentioned paper, which disappeared in ensuing debates: Hakala calls the URN as well as other discussed identification systems the "resolution services". This might seem as exorbitant pedantry to assert the difference between "resolution service" and "identification" or "identification system." But as this difference slowly faded away, our understanding of what we really mean by "URN:NBN" became more dissonant. Only after weakening Hakala´s stress on "resolution service", one may release a website with an URN:NBN generator without really harvesting generated identifiers and relevant metadata. Hakala´s paper is sufficiently clear in claiming that the pure "generating" of identifiers, i.e. generating numbers and letters without maintaining the register of the assigned identifiers and related metadata, is completely fruitless. Generating URN:NBN:CZ and not having a resolution system and a register at the same time is something like generating domain names without having IPs and DNS. To summarize, the URN:NBN infrastructure must be comprised of a resolution service module, administration module which also generates the numbers and a registry, the database containing all necessary information to enable the resolution of the digital documents.

After this starting point in 2001, the work on identifiers remained at the theoretical level for some time. Several Czech experts gained large theoretical knowledge of the existing identification systems for the digital world [2]. From 2007 onward, several libraries began to realize the needs for a practical solution of their specific problems with digital-born and digitized documents and started to use the Handle or other systems in practice.

NLCR made an attempt to establish cooperation in this area at the national level by setting up a Working Group for Persistent Identifiers in 2007. Unfortunately, the outcomes of this group's work resulted in a list of very general requirements and use cases for the URN:NBN system. If we really wish to end up with a national persistent identifier infrastructure based on the URN:NBN system, the time is ripe for a next step: getting back together and start working.

National bibliography identifier

The NLCR has very explicit goals resulting from its position in the library system. The reasons for using the URN:NBN in NLCR may be different from the ones in other institutions. In the following section we will try to explain in detail the ongoing challenges the NLCR has to cope with.

First of all, the large-scale digitizing project of the NLCR and Moravian State Library in Brno (thereinafter MLSB) coming up in the next few years anew calls for a solid identification system. If we are about to scan all Bohemical documents in the next 20 years, we need very careful reconsideration of our current identification systems and cataloging practices. The products of digitization have to be retrievable, the digital versions of the analog documents have to be bound with the catalog records of the analog documents and the catalog records must point to all relevant digital versions of the analog documents.

Well, this may not seem very problematic in an environment which strictly adheres to some standards, but the reality of the NLCR cataloging is quite loose. Existing catalog systems don't use FRBR; there are no identifiers of the works, and the Union Catalogue may seem to be far from unequivocal identification ideal. NLCR can sufficiently control the Czech National Bibliography catalog only, though even here the practice might fall short of ideal. The national bibliography identifier's (NBN) implementation might not lead to an absolutely perfect identification system, but we have yet decided to try it: the cataloging department of the NLCR will implement the national bibliography number identifier (also called as "CNB number"), which will be attached to all records in the Czech National Bibliography base. Since our mass digitization project will mainly focus on the Bohemical works, it seems to be the essential step towards long-term accessibility of the project's results. All digital representations should be bound to this identifier; the identifier of the basic "intellectual entity" (PREMIS). The basic "intellectual entity" will be an entity described by the Czech National Bibliography catalog.

Whatever the result of the nation-wide project of URN:NBN system's implementation is, we will try to use the URN:NBN:CZ to identify all the digital versions of analog documents described in the Czech National Bibliography catalog. This means that all Bohemical documents should be uniquely identified (does not matter if they have ISBN, ISSN or any other identifier, they should get the CNB number anyway) and all related digital representations (primarily those preserved in the NLCR and MSLB's long-term preservation digital repository system) will have URN:NBN:CZ. From our point of view, URN:NBN:CZ system registry should contain not only the URN, URL, but also the CNB number plus other descriptive and administrative metadata. The system should be able to search by CNB number, URN or other metadata fields returning all URN:NBN of the digital representations of the documents related to certain CNB number (and this means a catalog entity) and vice versa. The system should not only include identifiers of the user copies but also external identifiers linking to the preservation masters in the long-term preservation system of the NLCR and MSLB. We currently consider the possibility of using the existing RD:CZ database as a data source for this project, extending it to the web archiving results in the future.

National project for URN:NBN:CZ

Above in the text are the plans of the NLCR. Naturally, it would be possible to extend our plans in some way and ask other institutions to rigorously conform to the standards and concept of the URN:NBN, which we at some point (NLCR) present as final. But we think others should have their say. All those who call for URN:NBN:CZ identifier's implementation should be given a chance to say what they really need, what ideas they have and what they prefer at the administrative level. If the URN:NBN infrastructure has to work at the national level, consensus (for a long-term sustainability of the project's results) is critical.

Analyses of the existing systems showed there are several different implementations of the URN:NBN in the world. Basically, the difference is between those who use the URN:NBN to identify the "intellectual content" and those who simply identify the "computer files". Again, this might seem to be a pedantry, but we don't think it really is. The computer files can be identified by check sum hash functions like MD5, but the intellectual content has to be identified differently, since their intellectual content is what matters more than the digital file characteristics. This has a far reaching impact on the policy of the URN:NBN: What is a "new version" of an identified digital document? What are the significant changes that call for assigning a new identifier? Could changing the fonts of the PDF be characterized as the new version which would get a different URN:NBN identifier?

The second problem is how to administratively secure permanent accessibility of the identified objects. If the URN:NBN has to be persistent, the identified objects must be accessible in the future and permanently. This means, we need a chain of a safe long-term preservation repositories, not to mention a certification system, preservation standards, etc. Any other locations in the digital world are too volatile.

And even after this, can we really be sure that the objects we need to identify are of a permanent value? We will never be able to preserve the entire digital world. We need to select and assess the resources. The identifiers have to be here eternally, but how about the objects themselves? Selection criteria as well a policies are therefore more than needed.

Needles to say, any project implementing the URN:NBN has to clarify a number of administrative issues. Do we want to have a centralized system of resolution and administration or a decentralized one, will each institution have own resolution service and local register, would we only collect the data somewhere? If we opt for a centralized system, who and how will finance the system? Who will be responsible for maintenance, daily processes, etc? And if the solution needs to be sustainable it surely needs to be open and scalable. The architecture and data model has to be usable in the future in an altered environment, possibly for very different documents with different metadata. Will the system we design now really represent the solution we would choose in the future? In short, are we really sure that we need URN:NBN? Shouldn't we adopt the approach of the Library of Congress and choose for example PURL?

"A URN is (theoretically) a persistent identifier for a resource, independent of location or access method

…..

URNs never caught on because they tried to be too many things and never really nailed down which:

- *A persistent URL*
- *Location independent*
- *A resolution system*
- *A pure identifier*

Persistence and location independence came to be thought of more as social than technical problems. Other approaches were developed rather than formalizing the URN concept.

The proposed URN resolution system never was fully developed. And resolution is incompatible with the role of pure identifier." [3]

URN:NBN:CZ project: next steps

The NLCR will present the URN:NBN:CZ's concept to several interested institutions at the end of this month (October 2009). It is expected to proceed with a meeting in early November (2009), where we should discuss this model and assign responsibilities for project submission to VISK in January 2010. We hope the project to continue towards URN:NBN:CZ implementation in early 2011. The first steps should lead to 'polishing' the conceptualization, formulate the functional requirements, describe the use cases, the workflow, administration and communication interfaces of the system, delineate the data model, suggest integration with the catalogs or other systems. This should lead to the tender for programming-related works, which might or might not use RD.CZ base or any of the existing open-source software (e.g. the Italian URN:NBN software we have already tested). Besides, the administration and organization issues of the future system have to be settled. Much of the success of the URN:NBN:CZ's implementation depends on the administration and workflow, therefore the software solution should only conform to the needs of the participating institutions. All policies and responsibilities need to be very clearly defined.

Conclusion

We may only speculate about the very results of this nation-wide endeavor. As mentioned above, in any case, the NLCR, together with MSLB, will strive to use URN:NBN:CZ for identification of the Bohemical documents digitized in the upcoming mass digitization project. We have the responsibility to pass this data, the core of r culture heritage, to future generations. We will also try to preserve accessibility to the digital-born data from the web archiving project of the NLCR and MSLB. Although this might not be trivial, the URN:NBN should helps us here as well. How the URN:NBN:CZ will be applied in other parts of the librarian community or beyond it (and if it will be used at all), is not upon us to decide.

References

[1] HAKALA, Juha. 2003. Popis dokumentů a přístup k nim - nové výzvy. In KLOUČKOVÁ, Z. a MACHALOVÁ, L. (sest.). *Moderní informační a komunikační technologie v knihovnictví 2003 : sborník příspěvků.* Praha : Státní technická knihovna, 2003, s. 22-30. ISBN 80-86504-09-3.

[2] HUTAŘ, Jan; COUFAL, Libor. 2007. *Perspektivy trvalých identifikátorů v ČR* [online]. Praha : Národní knihovna ČR, 2007 [cit. 2009-10-27]. Dostupný z WWW: <http://www.ndk.cz/ochrana-digitalnich-dat/pid-1/PID/perspektivy-trvalych-identifikatoru-v-cr>.

BRATKOVÁ, Eva. 2007. *Síť trvalých identifikátorů informačních entit* [online]. Verze 1.0. Praha : Ústav informačních studií a knihovnictví FF UK v Praze, prosinec 2007 [cit. 2009-10-27]. Elektronické studijní texty ÚISK. Dostupný z WWW: <http://texty.jinonice.cuni.cz/>.

[3] Library of Congress. 2007. *URI resource pages : about URIs* [online]. Washington (DC, USA) : Library of Congress, 2007 [cit. 2009-10-27]. Dostupný z WWW: <http://www.loc.gov/standards/uri/about.html>.

About "INFO" URIs : Frequently Asked Questions [online]. 2006 , This version dated May 24, 2006 [cit. 2009-10-27]. Dostupný z WWW: <http://info-uri.info/registry/docs/misc/faq.html>.

[4]HAKALA, Juha. 2001. *Using national bibliography numbers as Uniform Resource Names* [online]. Helsinki : Helsinki University Library, 2001 [cit. 2009-10-27]. Dostupný z WWW: <http://www.ietf.org/rfc/rfc3188.txt>.