Archiving of the Slovak web space
E-journals preservation
report – september 2009

The project of University library in Bratislava – webdepozit – started with the early touches to this issue in september 2005. Thanks to the project with our Czech, Slovenian and Estonian partners, we were able to start with pilot webharvesting in 2006. The web-site for the project was established in August 2007 and the domain www.webarchiv.sk was registered. Later in 2008, the project and website were renamed to webdepozit (www.webdepozit.sk).
Our project is primarily oriented on the segment of scholarly and scientific continuing e-resources, in other words, the born-digital electronic journals and web sites, which are not systematicly stored and archived even though they belong to the cultural heritage of Slovak republic.

The quantitative and qualitative analysis of the „slovak web“ under top-level domain .sk, undertaken in May 2006, showed some first findings. In two years (from May 2006 to November 2008), the number of registered second-level domains almost doubled (to the total of 170.715 domains in November 2008). From the first pilot harvest in 2006, it was possible to gather some data showing the total and average size and number of the objects for selected web-pages. This was important for later harvester settings. Last year, we were able to organize hardware and (open-source) software for web archiving and in September 2008, we started the experimental selective harvesting of web-sites registered in the ISSN system. The whole-domain harvest started later in December 2008. Drawn from this experience, we will be able to formulate the criteria and strategy for the next practices.

Hardware and software

For the webdepozit project, in January 2008 University library in Bratislava acquired two data servers Sun Fire X4500, each driven by two dual-core AMD Opteron processors with 16 GB RAM and 48 TB of storage running the Solaris 10 operating system. For harvesting, we use danish NetarchiveSuite, which suits our needs after small modification – application uses secured interntet protocol and requests authentication of the user. For indexing, searching and access, we use Waybeck for searching via www address. We test wdsearch for full-text searching. Wdsearch is able to work with word flexion, this was not possible in NutchWax, software, which was used by another libraries and institutions, to access harvested materials. On the harvester-server, there run 2 jobs simultaneosly. Potential of the server is higher, however, if there are running 4 jobs, the response is poorer.

Suitable sources for selective harvesting and archiving

Webdepozit is oriented on the born-digital e-journals and web sites. These are either web sites with articles published in numbered issues (successively-issued e-journals), or updated web sites and databases (integrating resources). Every harvested source must be assigned with an ISSN number. This is one of the criteria for our selective harvesting. We made also pilot whole-domain (sk) harvesting, to test the possibilities and proper configuration of our software.

Selective harvesting – pilot project

In January 2008, our department made an inquiry, among the publishers, to map the situation on archiving of their e-journals and updating web sites. We contacted 176 publisher, received 94 responses. Most publishers answered, that they keep their own archives (over 91%). This was the most frequent reason, why the publishers were not intrested in cooperation with University library in Bratislava. They were affraid to let another institution to keep their data because of copyright and access to the data in archive in the future. Even, the web sites were free accessible without needed licences or registration. However, almost 70 % of asked publishers were interested in cooperation and building an external institutional deposit for their continuing resoures.

In contrast to the whole-domain harvesting, in selective harvesting, the web sites do not need to be under the national domain .sk. We can select also another, e.g. generic domains .com, .net., etc. See table 1 Web sites assigned with ISSN.

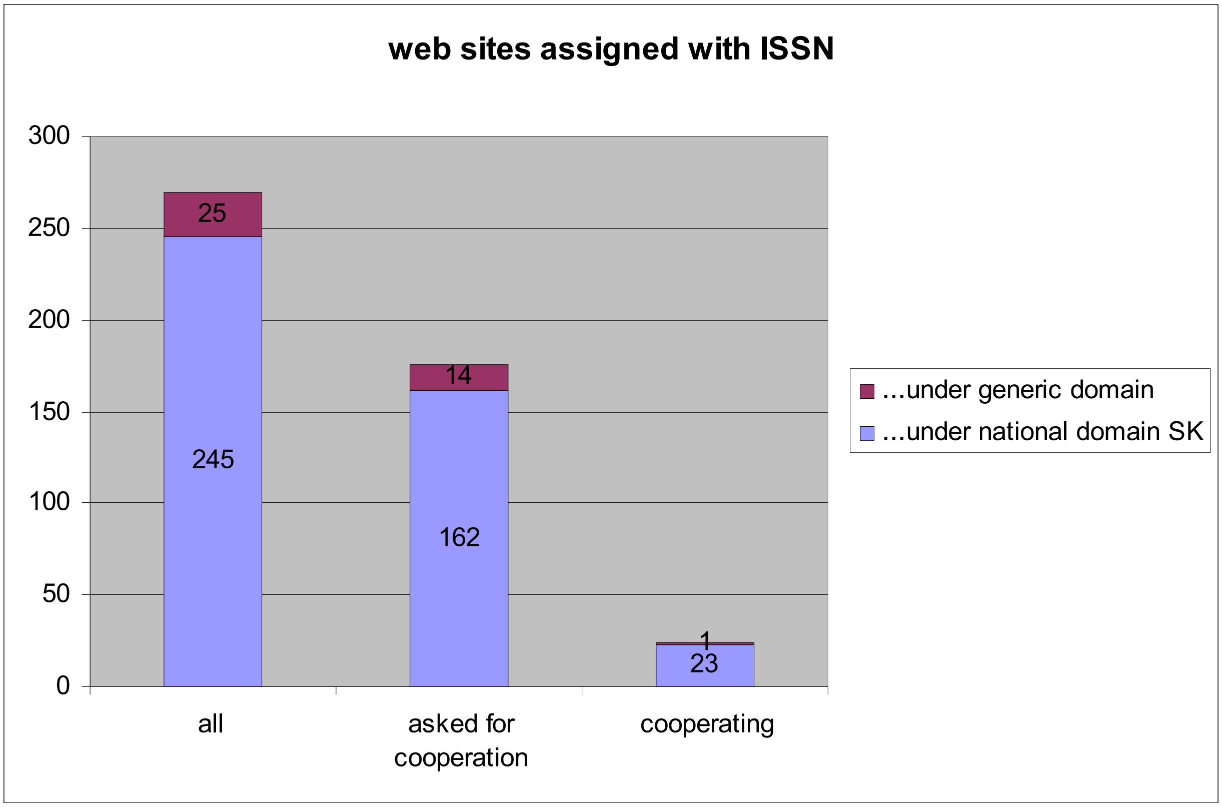


Table 1 Web sites assigned with ISSN

In the end-effect, we asked also new publishers, that applied for E-ISSN for their web-sites. We got harvesting-permission for 27 titles, see table 2 Pilot harvesting 2008.

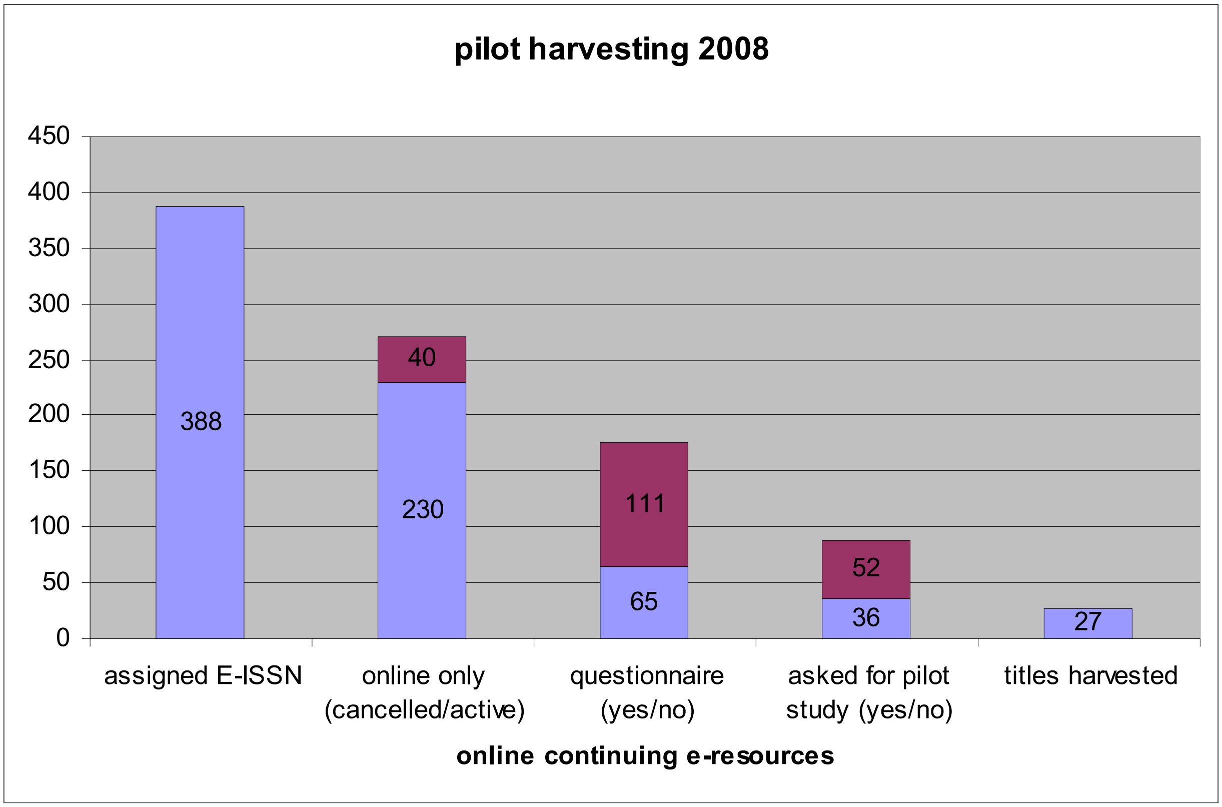


Table 2 Pilot harvesting 2008

These 27 harvested web sites can be divided into two big groups – integrating resources (updating web sites and databases in form of portals) and successively issued (one or more texts from one or more authors in form of classical e-journals). See table 3 Used format in harvested continuing resources.

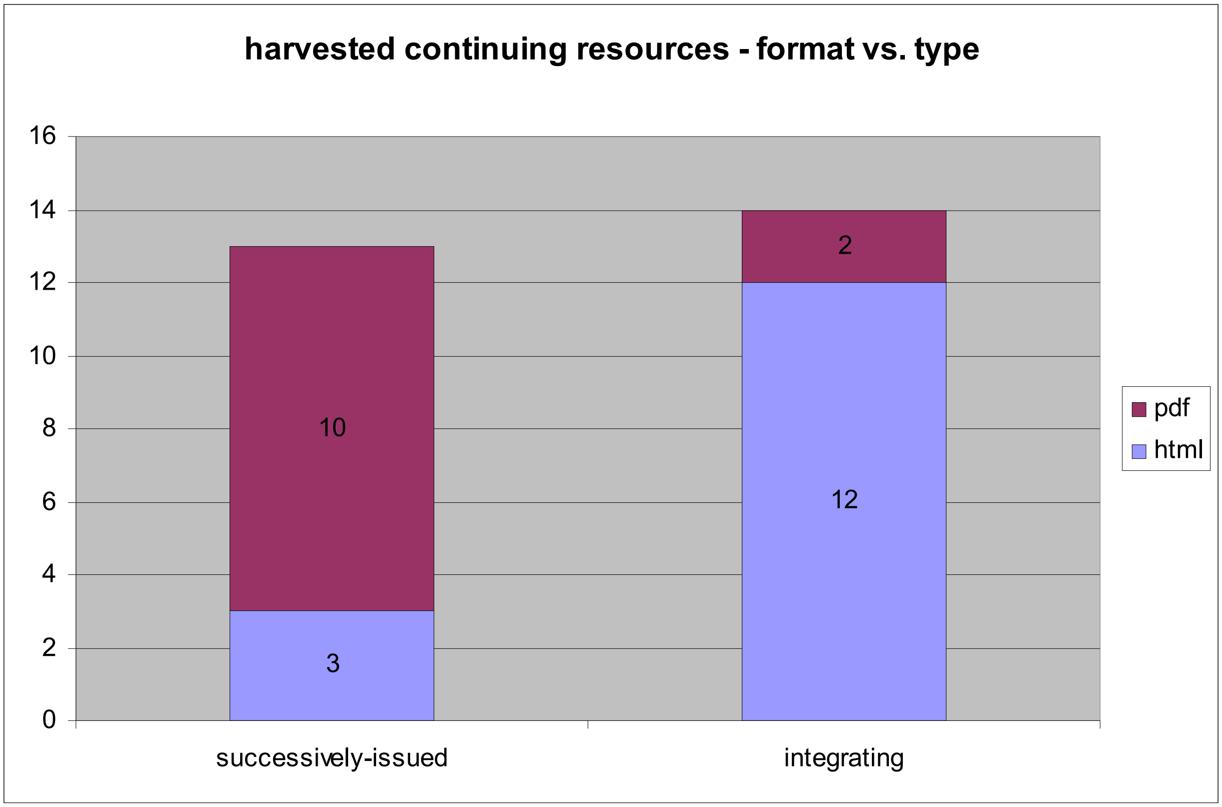
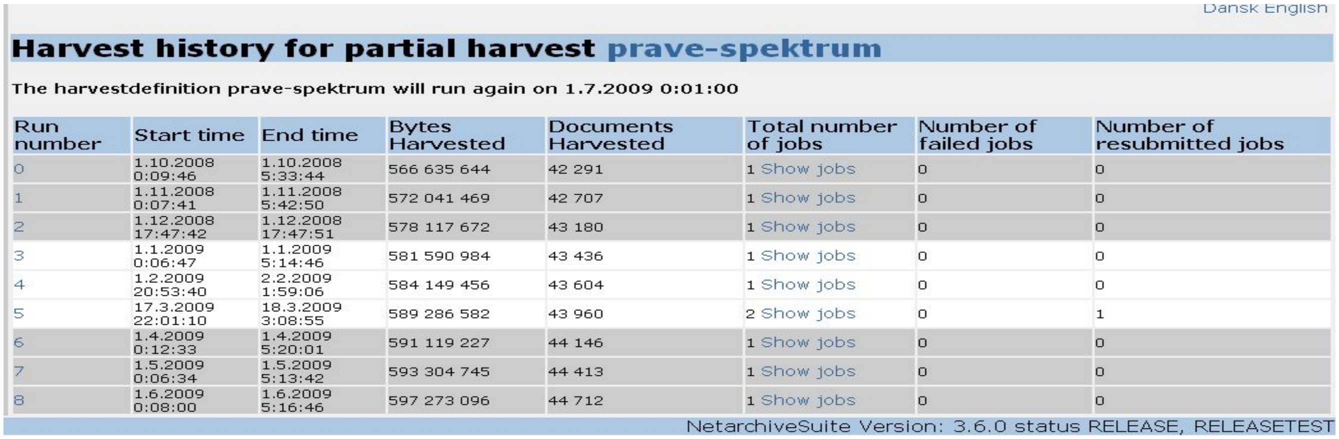


Table 3 Used format in harvested continuing resources

During the pilot harvesting, we decided to use rss for article harvesting from the portal sites, because there is a lot of content, we are not interested in. The positive point is, that we can determine scope of the harvest quite exactly and take only pages/seeds with articles. Thanks to this solution, the amount of harvested data decreased from 56 MB to 9KB. See Picture 1 and 2



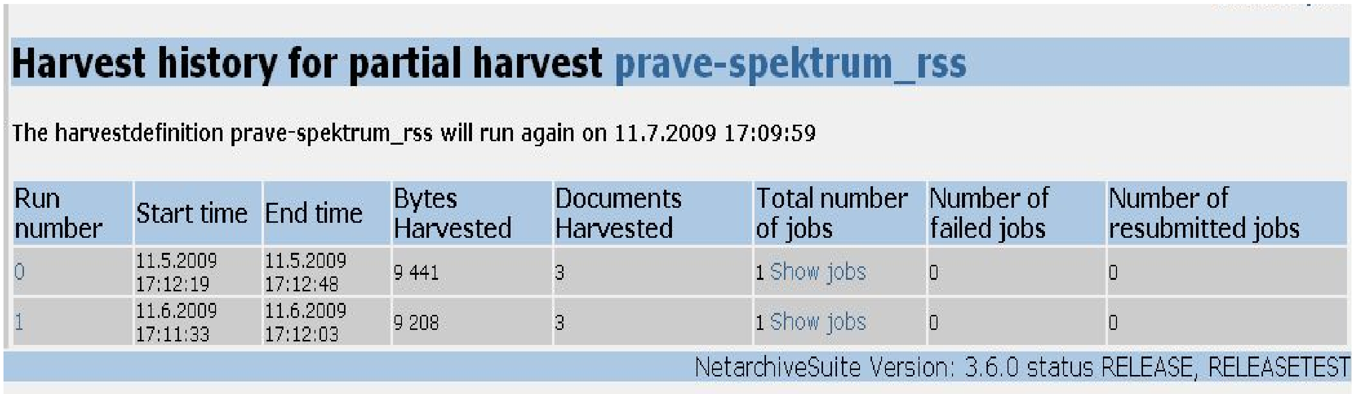
Harvest history for partial harvest prave-spektrum

The harvestdefinition prave-spektrum will run again on 1.7.2009 0:01:00

Run number	Start time	End time	Bytes Harvested	Documents Harvested	Total number of jobs	Number of failed jobs	Number of resubmitted jobs
0	1.10.2008 0:09:46	1.10.2008 5:33:44	566 635 644	42 291	1 Show jobs	0	0
1	1.11.2008 0:07:41	1.11.2008 5:42:50	572 041 469	42 707	1 Show jobs	0	0
2	1.12.2008 17:47:42	1.12.2008 17:47:51	578 117 672	43 180	1 Show jobs	0	0
3	1.1.2009 0:06:47	1.1.2009 5:14:46	581 590 984	43 436	1 Show jobs	0	0
4	1.2.2009 20:53:40	2.2.2009 1:59:06	584 149 456	43 604	1 Show jobs	0	0
5	17.3.2009 22:01:10	18.3.2009 3:08:55	589 286 582	43 960	2 Show jobs	0	1
6	1.4.2009 0:12:33	1.4.2009 5:20:01	591 119 227	44 146	1 Show jobs	0	0
7	1.5.2009 0:06:34	1.5.2009 5:13:42	593 304 745	44 413	1 Show jobs	0	0
8	1.6.2009 0:08:00	1.6.2009 5:16:46	597 273 096	44 712	1 Show jobs	0	0

NetarchiveSuite Version: 3.6.0 status RELEASE, RELEASETEST

Picture 1 prave-spektrum – seed is the homepage www



Harvest history for partial harvest prave-spektrum_rss

The harvestdefinition prave-spektrum_rss will run again on 11.7.2009 17:09:59

Run number	Start time	End time	Bytes Harvested	Documents Harvested	Total number of jobs	Number of failed jobs	Number of resubmitted jobs
0	11.5.2009 17:12:19	11.5.2009 17:12:48	9 441	3	1 Show jobs	0	0
1	11.6.2009 17:11:33	11.6.2009 17:12:03	9 208	3	1 Show jobs	0	0

NetarchiveSuite Version: 3.6.0 status RELEASE, RELEASETEST

Picture 2 prave-spektrum – seed is the rss-channel

Hovewer, rss channels do not work as OAI-PMH, that means you can not access archive or older articles via rss. For number of seeds by methods used for harvesting, see table 4 Used harvest-method in harvested continuing resources.

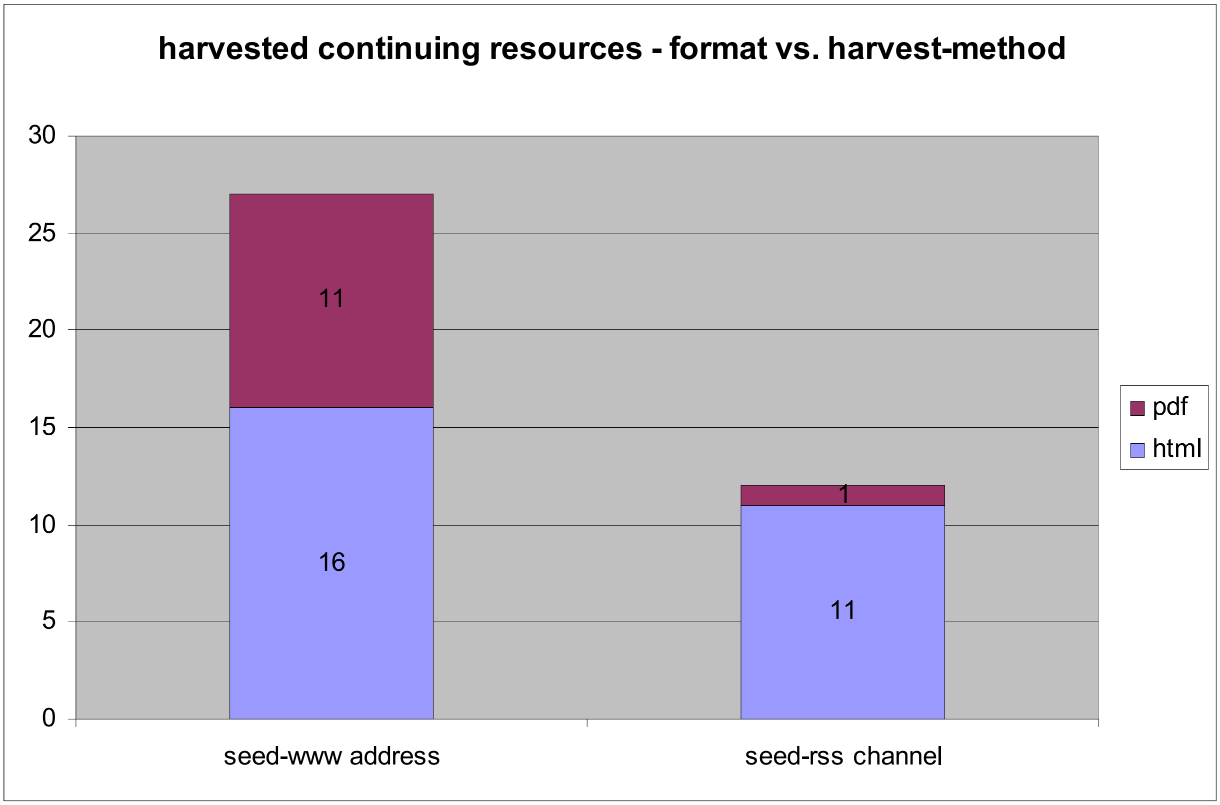


Table 4 Used harvest-method in harvested continuing resources

Next steps and plans

We continue with harvesting and we started to index the harvested content. As mentioned before, we will have two possibilities to access the content – in Wayback via www address and in wdserach via key words. However, we have to solve the problem with indexing of the pdf content. From the testing harvests, only 3 publishers used Dublin core metadata for published articles, even these contained only meta name title or author, nobody used meta name description, (if yes, there was only the title of the article written).

Later, for the routine process, we plan to put new question to the ISSN assignment request for online resources. The publisher will choose, if he agrees with the cooperation on building the webdepozit. In an agreement (we have to draw up, because there is no legal deposit for online resources in Slovakia), the publisher will give his permisson to make his resource accessible via webdepozit (under given restrictions). The long-term preservation of archived resources is challenge, we have to face, before we proceed to routine harvesting and agreements with publishers.

Katarína Kovačičová
e-librarian
University library in Bratislava
www.ulib.sk
www.webdepozit.sk