

Webarchiv

Památník českého internetu, [více](#)

Proč archivujeme web?

Jak ho archivujeme?

Co v archivu máme?

Co nás trápí?

Co plánujeme?

Rudolf.Kreibich@nkp.cz

technická realizace archivace webu

Proč archivujeme web?



Národní filmový archiv



O nás

Aktuální

Oddělení

Ponrepo

Nabízíme

Vyhledávání

Vážení diváci,

přehlídka Ruská sezóna 2006 - Současný ruský film (Naši, Řidič pro Věru, Nejen chlebem, Můj nevlastní bratr Frankenstein, Nankinská krajina, Slunce) byla z technických důvodů přesunuta na listopad 2006. V náhradním programu uvidíte:

Francouzský režisér a herec Gérard Oury (1919 - 2006).



Využití sbírek NFA Ponrepo On-line katalog knihovny Digitální knihovna Filmový přehled Časopis Iluminace

Hledaný výraz OK

NÁRODNÍ FILMOVÝ ARCHIV

Vítáme Vás na stránkách NÁRODNÍHO FILMOVÉHO ARCHIVU, které mají od prosince 2008 novou podobu. Doufáme, že zde pohodlně najdete potřebné informace, a že se Vám náš nový web bude líbit! Vaše náměty a dotazy uvítáme na adrese nfa@nfa.cz.



NFA je na základě §80 zákona č. 499/2004 sb. akreditovaným specializovaným archivem a jeho posláním je shromažďování, ochrana, vědecké zpracování a využívání audiovizuálních archiválií dokumentujících národní filmovou produkci, vznik a vývoj filmového umění, život českého národa a významné světové události.

V současné době NFA pečuje o více než 150 milionů metrů filmu, 500 tisíc fotografií, 40 tisíc plakátů, 134 tisíc propagačních materiálů, rozsáhlý archiv českého filmu a filmovou knihovnu.

[více >>](#)

Rozšířený filmový sbírek v Čechách

AKTUALITY

19. 4. Hudební doprovody k němým filmům Alfreda Hitchcocka

Klavíristé Karel Loula, Petr Wajsar a Vilém Endler doprovází Hitchcockovy němé filmy Světový šampión (20.4.), Na šikmé ploše (23.4.) a A neuved' nás v pokušení (28.4.).

[více >>](#)

8. 4. Německé publikace o filmu v prodeji v Ponrepo

Lednová přehlídka německých filmů CineFest (Ponrepo, 7.1. - 13.1.2009) přinesla vedle zajímavých filmů také možnost zakoupit si ojedinělé publikace o filmu.

[více >>](#)

1. 4. Volná pozice

Národní filmový archiv přijme pracovníka na plný úvazek.

[více >>](#)

4. 12. Ponrepo má vlastní web!

Vážení diváci, navštivte nové

Homepage

O nás

Kontakty

Oddělení NFA

Využití sbírek NFA

Filmy k distribuci

Knihovna

Publikace

Projekty

Filmová ročenka

BIO PONREPO



Nejbližší filmy

21. 4. 2009 17:30

Cirkus

21. 4. 2009 20:00

Kráska z Marsu



22. 4. 2009 17:30

Skřiváncí ticho



NFA

FB →

Vyhledávání → EN →

Aktuality

02/01

Vyhlášíme veřejnou zakázku Digitalizace a digitální restaurování filmu Staré pověsti české

Digitalizace a digitální restaurování bude probíhat v rozlišení až 4K s ohledem na stav zdrojových materiálů. Výstupem bude DCP určené ke kinodistribuci a dlouhodobé archivaci, master pro výrobu DVD a BD a podrobná dokumentace celého procesu. Lhůta pro podání nabídek je 13. ledna do 15:00. Veřejnou zakázku najdete na profilu zadavatele GORDION.

[→ více](#)

01/01

Leden ve znamení Visegrádu

Lednové nedělní projekce Ponrepo dětem jsou věnovány krátkým animovaným filmům zemí Visegrádu, které se do kmenového kina NFA Ponrepo dostanou díky mezinárodnímu projektu Animace v akci! Začínáme každou neděli jako vždy ve tři odpoledne a projekce budou doprovázeny tvůrčími dílnami. Spolupráce Národního filmového archivu s kolegy ze Slovenska, Polska a Maďarska umožní...

[→ více](#)

16/12

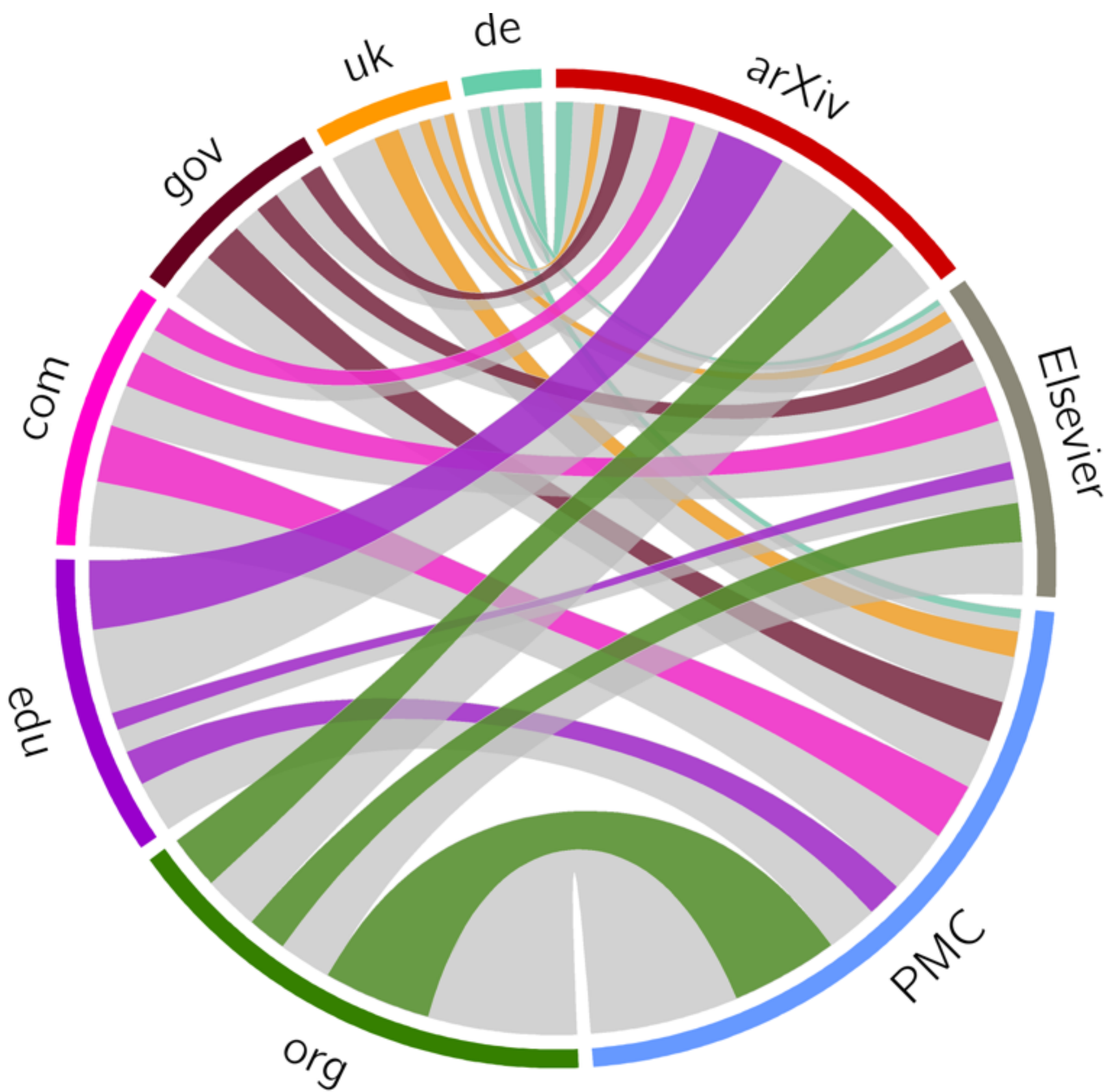
Pracovní příležitost v Knihovně NFA

Knihovna Národního filmového archivu hledá kandidátku/kandidáta na pozici knihovníka. Národní filmový archiv (www.nfa.cz) je státní příspěvkovou organizací Ministerstva kultury ČR. Předpokládáný nástup: únor 2015 Pracoviště: Hradištko-Brunšov Příjem přihlášek: do 20. ledna 2015 Osobní pohovory: 27. ledna 2015 na pracovišti v Praze Práce v rozsahu plného úvazku s možností zkráceného úvazku Pracovní náplň: správa depozitáře knihovního fondu samostatné organizování knihovního fondu spolupráce...

[→ více](#)

“... více jak 70% URL v Harvard Law Review a 50% URL v nálezech nejvyššího soudu Spojených států amerických, neodkazuje k původnímu webovému zdroji. “

Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. Jonathan Zittrain, Kendra Albert a Lawrence Lessig. Legal Information Management / Volume 14 / Issue 02 / June 2014, pp 88-99, DOI: <http://dx.doi.org/10.1017/S1472669614000255>, Published online: 12 June 2014



404 Not Found

The 404 (Not Found) status code indicates that the origin server did not find a current representation for the target resource or is not willing to disclose that one exists. A 404 status code does not indicate whether this lack of representation is temporary or [permanent; the 410 \(Gone\) status code is preferred](#) over 404 if the origin server knows, presumably through some configurable means, that the condition is likely to be permanent.

A 404 response is cacheable by default; i.e., unless otherwise indicated by the method definition or explicit cache controls (see [Section 4.2.2 of \[RFC7234\]](#)).

+

url *urn*

*Bude možné studovat naše století bez
webových archivů?*

miluj výzkumníka svého



FIGURE 5

Subquestion 3: What types of born digital news content do researchers use in their research, compared with the types of born digital news content they consider important to preserve?

Born digital news content researchers use	Percentage	What researchers want preserved (1=unimportant, 3=somewhat important, 5=important)
Born digital articles	59%	3.9
News blogs	38%	3.0
Online polls	21%	2.9
Info graphics	28%	3.3
Tweets	3%	2.2
Facebook posts	5%	2.2
Streaming videos	14%	3.1
Video transcripts	12%	3.4
Online ads	0%	1.9
Don't use born digital content	29%	
Not sure	9%	

Jak archivujeme web?

Heritrix / *OpenWayback*
sklizení / zpřístupnění

Otevřený software
Mezinárodní komunita

Měsíční výběrové sklizně

Občasné tématické sklizně

Roční sklizně domény cz

MORE IS MORE.

LESS IS LESS.

FIGHT MONOCULTURES.

181
803
399

Co máme v archivu?

~ 228 TB

~ 6 miliard digitálních objektů / URL

~1,2 milónu domén druhého řádu

Webarchiv podle oborů

Přehled nasmlouvaných webů dle oborového třídění:

[Vše](#) 4738 / [Antropologie, etnografie](#) 133 / [Beletrie](#) 29 / [Biologické vědy](#) 200 / [Divadlo, film, tanec](#) 153 /
[Ekonomické vědy, obchod](#) 277 / [Filozofie a náboženství](#) 222 / [Fyzika a příbuzné vědy](#) 86 /
[Geografie, Geologie, Vědy o Zemi](#) 366 / [Historie a pomocné historické vědy, Biografické studie](#) 237 / [Hudba](#) 146 /
[Chemie, Krystalografie, Mineralogické vědy](#) 45 / [Jazyk, lingvistika, literární věda](#) 173 /
[Knihovnictví, informatika, všeobecné, referenční literatura](#) 253 / [Lékařství](#) 267 / [Literatura pro děti a mládež](#) 1 /
[Matematika](#) 42 / [Politické vědy \(Politologie, politika, veřejná správa, vojenství\)](#) 315 / [Právo](#) 142 / [Psychologie](#) 83 /
[Sociologie](#) 278 / [Technika, technologie, inženýrství](#) 264 / [Tělesná výchova a sport, Rekreace](#) 207 /
[Umění, architektura](#) 316 / [Výchova a vzdělávání](#) 188 / [Výpočetní technika](#) 143 / [Zemědělství](#) 172

Historie a pomocné historické vědy. Biografické studie

[Vše](#) 237 / [Archeologie](#) 24 / [Biografie](#) 8 / [Dějiny \(obecně\)](#) 21 / [Dějiny civilizace, Kulturní dějiny](#) 10 /
[Dějiny Česka a Slovenska](#) 24 / [Dějiny Evropy](#) 3 / [Dějiny středověku a novověku](#) 2 /
[Dějiny zemí starověkého světa](#) 3 / [Dějiny zemí střední Evropy](#) 5 / [Dějiny zemí východní Evropy](#) 1 /
[Genealogie, Heraldika, Šlechta, Vlajky](#) 16 / [Geografie, Biografické studie, Dějiny](#) 1 /
[Historická věda, Pomocné vědy historické, Archivnictví](#) 27 / [Kulturní památky historického období obecně](#) 86 /
[Pravěk](#) 2 / [Studium oblastí a lokalit obecně](#) 1 / [Světové dějiny](#) 3

Zobrazení: [vizuální](#), [textové](#)



[Akce Památky : pro kulturní dědictví a zábavu](#)



[Analýza 17. listopadu 1989](#)



[Antikomplex](#)



[Antologie ideologických textů : Ústav pro studium totalitních režimů](#)

Co nás trápí?

méně než 1 % je volně přístupné

=

~ 4000 webů z 1,2 miliónu webů

JavaScript



Open access časopisy Vydavatelství FF UK

[Více...](#)

Časopis pro moderní filologii (18)



Časopis pro moderní filologii je recenzovaný neimpaktovaný jazykovědný časopis. [Více...](#)

```
...tion_id=3227'
<!-- www_load_file_ng : /exlibris/dtl/u3_1/dtle/www_r_eng/doctype-declaration -->
<!-- FILE-STARTS : /exlibris/dtl/u3_1/dtle/www_r_eng/doctype-declaration -->
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">

<!-- FILE-ENDS : /exlibris/dtl/u3_1/dtle/www_r_eng/doctype-declaration -->

<!-- www_load_file_ng : /exlibris/dtl/u3_1/dtle/www_r_eng/pds-sso -->
<!-- FILE-STARTS : /exlibris/dtl/u3_1/dtle/www_r_eng/pds-sso -->

<html>
<head>
  <title>PDS SS0</title>
<script>
<!--

var debug =0 ;

function SSOPds() {
  var url = 'https://digitool.is.cuni.cz:443/pds?func=sso&calling_system
=digitool&pds_con_lng=&con_lng=&url=http://digitool.is.cuni.cz/R/U3H8
D6T3CT8UGHLHIY62T4RJX1333QX6HCKHCCQAKRDY67TEAM-01642?func=collections&collec
tion_id=3227';
  if (debug == 1) alert (url);
  window.location = url;
}
```

```
10 <HTML>
11 <HEAD>
12 <TITLE>DigiTool - Digitální sbírky</TITLE>
13 <!-- FILE-STARTS : /exlibris/dtl/u3_1/dtle/www_r_cze/header-include -->
14 <META HTTP-EQUIV="Expires" CONTENT="Tue, 20 Aug 1996 14:25:27 GMT">
15 <META HTTP-EQUIV="REFRESH" CONTENT="2000; URL=http://digitool.is.cuni.cz/R/8BLMTC8XHIX1TQSL97XDJBGAMEAMGFYGCTIJ4998JDBR9PUAFE-00103?
func=exit">
16 <META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=UTF-8">
17 <link rel="stylesheet" type="text/css" href="http://digitool.is.cuni.cz/rd.css">
18
19 <!-- FILE-ENDS : /exlibris/dtl/u3_1/dtle/www_r_cze/header-include -->
20 <script src="http://digitool.is.cuni.cz/exlibris/dtl/u3_1/dtle/www_r_cze/js/popup.js" type="text/javascript"></script>
21 <script src="http://digitool.is.cuni.cz/exlibris/dtl/u3_1/dtle/www_r_cze/js/scripts.js" type="text/javascript"></script>
22 </HEAD>
23
24 <BODY>
25
26 <!-- END_SECTION --><!-- head -->
27
28 <!-- www_load_file_ng : /exlibris/dtl/u3_1/dtle/www_r_cze/collections-head -->
29 <!-- START_SECTION --><!-- banner -->
30 <!-- FILE-STARTS : /exlibris/dtl/u3_1/dtle/www_r_cze/banner -->
31 <TABLE CELLPADDING="0" CELSPACING="0" BORDER="0" width="100%">
32   <TR>
33     <TD>
34       <!-- Banner starts -->
35       <TABLE CELLPADDING="0" CELSPACING="0" BORDER="0" width="100%">
36         <TR>
37           <TD><IMG SRC="http://digitool.is.cuni.cz/exlibris/dtl/u3_1/dtle/www_r_cze/icon/header_left.gif" WIDTH="111" HEIGHT="100"
BORDER="0" Title="DigiTool Správa digitálních dokumentů" ALT="DigiTool Správa digitálních dokumentů"></TD>
38           <TD class="banner" width="100%"><IMG SRC="http://digitool.is.cuni.cz/exlibris/dtl/u3_1/dtle/www_r_cze/icon/pixel.gif" WIDTH="1"
HEIGHT="100" BORDER="0" Title="" ALT=""></TD>
39           <TD align="right"><IMG SRC="http://digitool.is.cuni.cz/exlibris/dtl/u3_1/dtle/www_r_cze/icon/header_right.gif" WIDTH="495"
HEIGHT="100" BORDER="0" Title="" ALT=""></TD>
```

228 TB při čtení 80 MB/s = 32 dní

Co plánujeme?

metadata

☒

ARJ archive data, v6, slash-switched, original name: SEAS95.ARJ, os: MS-DOS

Audio/Video Interleaved Format

Conflicted

Exchangeable Image File Format

Extensible Markup Language

Flash Video (FLV)

Graphics Interchange Format

Hypertext Markup Language

JPEG File Interchange Format

LHa (2.x) archive data [lh5]

MPEG 1/2 Audio Layer 3

Microsoft Excel

Microsoft Powerpoint Presentation

Microsoft Word Document

OpenDocument Text

PCX

PDF/A

PDF/X

Plain text

Portable Document Format

Portable Network Graphics

PostScript

Scalable Vector Graphics (SVG)

Tagged Image File Format

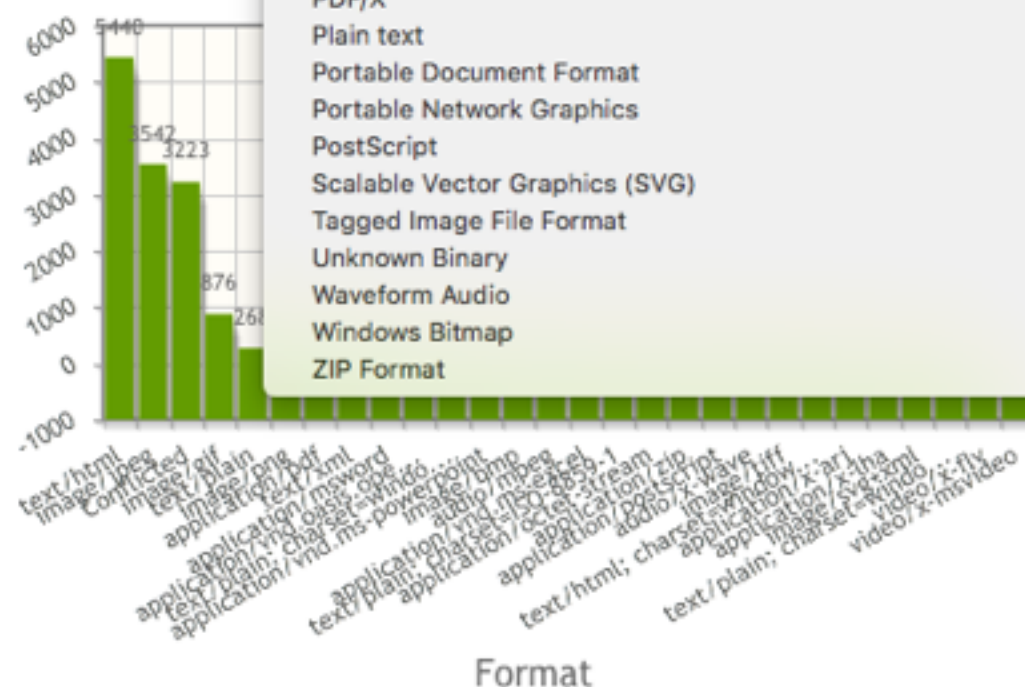
Unknown Binary

Waveform Audio

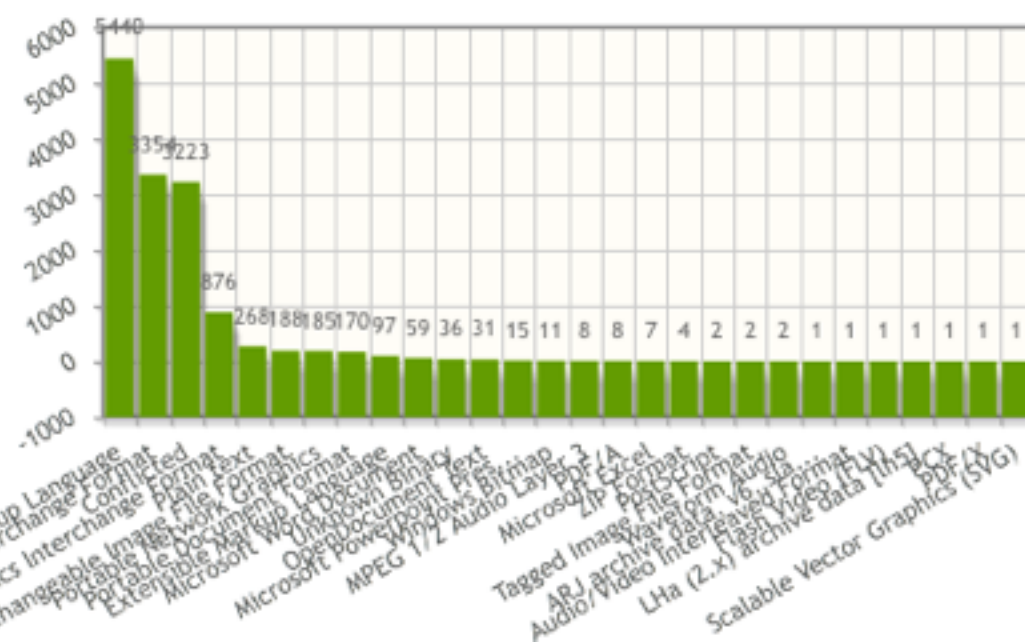
Windows Bitmap

ZIP Format

2001_20:



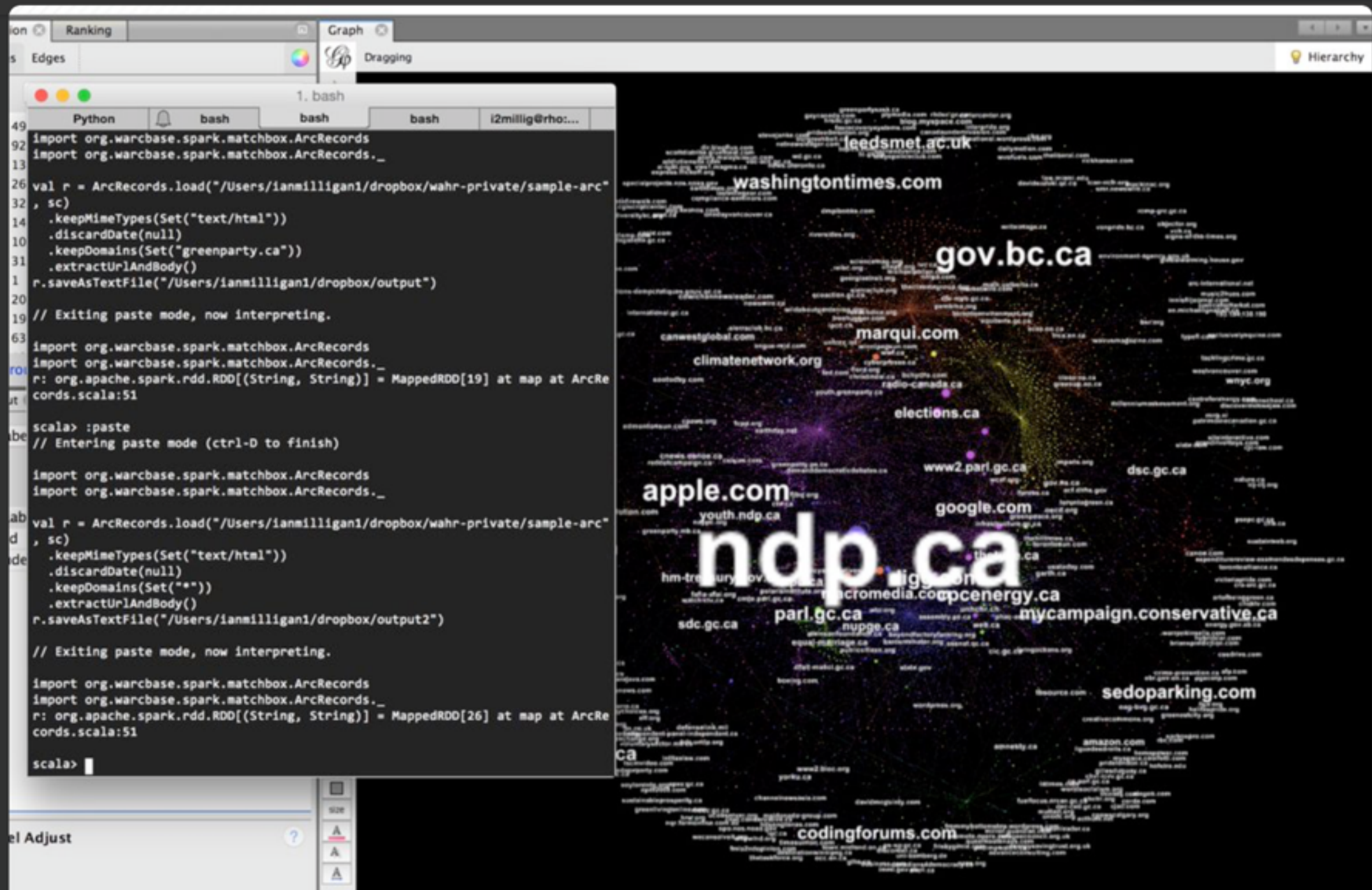
Format



Add Diagram

Size Stats

Count : 13993 objects
Size : 889,3 MB
Min : 1 B
Max : 180,63 MB
Average : 0,06 MB
SD : 1,66 MB
Var : 2,77 MB



Ian Milligan @ianmilligan1 · Oct 30

Last minute #webArchiving workshop tweaking... 2 hours of web scraping, text & network analysis. Hope it all works!



1



4





BATCH
(MapReduce)

INTERACTIVE
(Tez)

ONLINE
(HBase)

STREAMING
(Storm, S4,...)

GRAPH
(Giraph)

IN-MEMORY
(Spark)

HPC MPI
(OpenMPI)

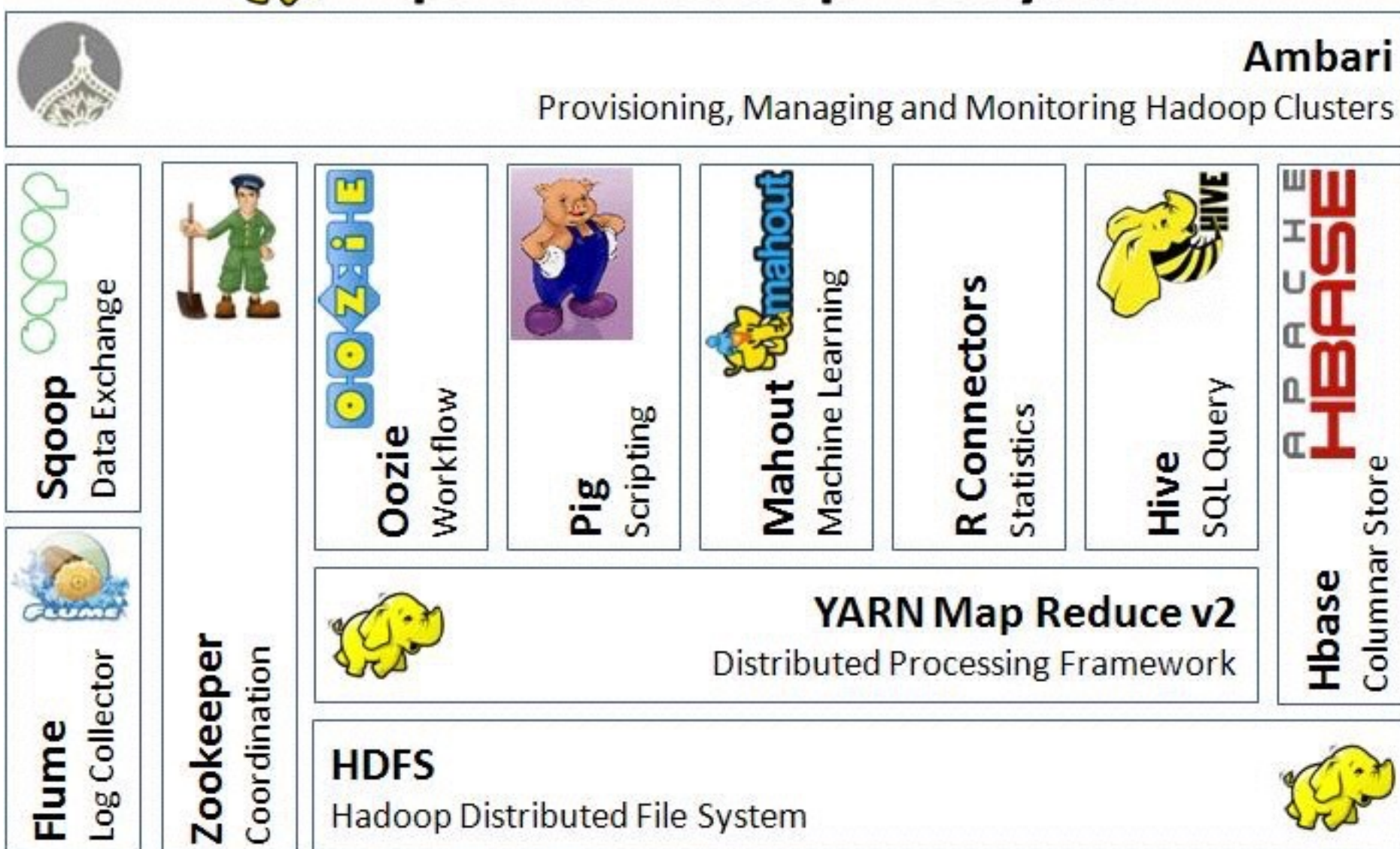
OTHER
(Search)
(Weave...)

YARN (Cluster Resource Management)

HDFS2 (Redundant, Reliable Storage)



Apache Hadoop Ecosystem





Oddělení archivace webu / ODIF / Národní knihovna ČR

Vedoucí: Jaroslav Kvasnica

Zástupce: Barbora Rudišínová

Technické zajištění: Rudolf.Kreibich@nkp.cz

Kurátoři: Markéta Hrdličková, Pavla Kupcová

webarchiv.cz

facebook.com/webarchivcr

slideshare.net/webarchivCZ

github.com/webarchivcz