Otevřená data a kam s nimi?

Tereza Simandlová | @kliste
tereza.simandlova@techlib.cz

Open Access aneb Open your mind! 2012, NTK, 24. 10. 2012

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Technical Library


Data

Co jsou data?

„…vyjádření skutečností formálním způsobem tak, aby je bylo možno přenášet nebo zpracovat (např. počítačem).“ (Wikipedie)

„Reprezentace informací vhodně formalizovaná pro komunikaci, interpretaci a zpracování lidmi a automaty. Data mohou být reprezentována libovolnými řetězci znaků (čísel, příkazů, vět) uloženými na informačním nosiči. Data nemají zpravidla význam sama o sobě, ale teprve jsou-li ***pochopena, interpretována, komunikována*** *a* ***využita*** *člověkem nebo počítačem, stávají se smysluplnými informacemi.“* (TDKIV)

Index skutečnosti

Otevřená data

„Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control." (Wikipedia)

otevřená vědecká data - Open Research Data

otevřená vládní data - Open Government Data

4 Challenges for Open Data

1. Free is not always open
2. Open is not always free
3. Analysis is not always easy
4. Open data is not always good

Zdroj: Tennison, Jeni. 2012. *Four Challenges for Open Data*. Available from: http://www.theodi.org/blog/four-challenges-open-data

Otevřená vědecká data

"*Open science data is a type of open data focused on publishing observations and results of scientific activities available for anyone to analyze and reuse.*" (Wikipedia)

Otevřená data jsou součástí Open Science (otevřené vědy)

The Panton Principles for Open Data (2010)

Panton Principles for OD

*„For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made **open**.“*

1. When publishing data make an explicit and robust statement of your wishes.
2. Use a recognized waiver or license that is appropriate for data.
3. If you want your data to be effectively used and added to by others it should be open as defined by the Open Knowledge/Data Definition – in particular non-commercial and other restrictive clauses should not be used.
4. Explicit dedication of data underlying published science into the public domain via PDDL or CCZero is strongly recommended and ensures compliance with both the Science Commons Protocol for Implementing Open Access Data and the Open Knowledge/Data Definition.

Panton Principles, Principles for open data in science. Murray-Rust, Peter; Neylon, Cameron; Pollock, Rufus; Wilbanks, John; (19 Feb 2010). Retrieved [insert date] from http://pantonprinciples.org/

Zvýšení transparentnosti ve vědě

Zvýšení efektivity
časová a finanční
úspora

Problémy a nevýhody

legislativní

technické

lidský faktor

Datové repozitáře

Úložiště pro data (datasety)
Institucionální / oborové / multioborové

Výhody (pro vědce):

- Dostupnost
- Viditelnost
- Dlouhodobá ochrana

Kde najít svůj repo?

OAD - Data repositories

Databib

DataCite - repolist

re3data.org – právě probíhající projekt, cílem je vytvořit globální registr repozitářů pro vědecká data

Edinburgh DataShare

Search Edinburgh DataShare

Go

Search Edinburgh DataShare
This Collection

Advanced Search

Browse

Edinburgh DataShare
Research Communities
Date Issued
Data Creators
Titles
Subjects
This Collection
Date Issued
Data Creators
Titles
Subjects

My Account

Login
Register

Statistics

View Statistics

children_Agar

Download Dataset

Show full item record

Title:	children_Agar
Data Creator:	Remijsen, Bert; Impey, Angela; Ajuet Deng, Elizabeth Achol; Deng Yak, Simon Yak; Ayuel Ring, Peter Malek; Penn de Ngong, John; Reid, Tatiana; Ladd, D. Robert; Meyerhoff, Miriam
Date Available:	2012-10-16
Citation:	Remijsen, Bert; Impey, Angela; Ajuet Deng, Elizabeth Achol; Deng Yak, Simon Yak; Ayuel Ring, Peter Malek; Penn de Ngong, John; Reid, Tatiana; Ladd, D. Robert; Meyerhoff, Miriam. (2012). children_Agar, 2009-2012 [sound]. University of Edinburgh, School of Philosophy, Psychology and Language Sciences.
Permanent Identifier:	http://hdl.handle.net/10283/217
Dataset Description (abstract):	The songs in this collection were recorded and annotated as part of the project 'Metre and Melody in Dinka Speech and Song', a project carried out by researchers from the University of Edinburgh and the School of Oriental and African Studies in London, and funded by the UK Arts and Humanities Research Council as part of their 'Beyond Text' programme. The project aimed to understand the interplay between traditional Dinka musical forms and the Dinka language (which distinguishes words not just by different consonants and vowels but also by means of rhythm, pitch and voice quality), and to learn more about the way the song tradition responded to the disruptions of the long Sudanese civil war. In this context, we aimed to record a large collection of Dinka songs for preservation in a long-term sound archive. This collection is the result of that effort. It presents song material from 36 Dinka singers and groups of singers. Further details can be found in the readme file. The collection is accompanied by an index, which is explained in the readme file.

Files	Description	Size	Format	View
ACollectionofDinkaSongs_index.xlsx		68.96Kb	Microsoft Excel	Download
ACollectionofDinkaSongs_readme.docx		46.49Kb	Microsoft Word	Download
ACollectionofDinkaSongs_readme.pdf		165.2Kb	PDF	Download
IMG_1243.JPG		464.3Kb	JPEG image	
IMG_1244.JPG		494.2Kb	JPEG image	
IMG_1245.JPG		405.3Kb	JPEG image	
IMG_1246.JPG		427.1Kb	JPEG image	
IMG_1247.JPG		426.1Kb	JPEG image	

http://datashare.is.ed.ac.uk/

TreeBASE *A Database of Phylogenetic Knowledge*

- Search TreeBASE
- Submission Tutorial
- Submit
- About
  - Overview
  - Technology
  - People
  - Partnerships
  - References
- NSF Data Management
- Data Access
- Journals
- Contact

Welcome to TreeBASE

TreeBASE is a repository of phylogenetic information, specifically user-submitted phylogenetic trees and the data used to generate them. TreeBASE accepts all types of phylogenetic data (e.g., trees of species, trees of populations, trees of genes) representing all biotic taxa. Data in TreeBASE are exposed to the public if they are used in a publication that is in press or published in a peer-reviewed scientific journal, book, conference proceedings, or thesis. Data used in publications that are in preparation or in review can be submitted to TreeBASE but are only available to the authors, publication editors, or reviewers using a special access code. TreeBASE is produced and governed by the **The Phyloinformatics Research Foundation, Inc.**

Some recent additions:

- **Molecular phylogeny of microhylid frogs (Anura: Microhylidae) with emphasis on relationships among New World genera**
  October 20, 2012
- **Novel and highly diverse fungal endophytes in soybean revealed by the consortium of two different techniques**
  October 20, 2012
- **Phylogeny of Escallonia (Escalloniaceae) based on plastid DNA sequence data**
  October 19, 2012
- **The invasive medicinal plant Piper aduncum supporting a new Pseudocercospora species**
  October 16, 2012
- **Miocene divergence, cryptic lineages, and contrasting distribution patterns in common camoflauge lichen fungi (Parmeliaceae, Ascomycota)**
  October 16, 2012
- **When Naked became Armored: An Eight-Gene Phylogeny Reveals Monophyletic Origin of Theca in Dinoflagellates.**
  October 16, 2012
- **Neogene-Dominated Diversification in Neotropical Montane Lichens: Dating Divergence Events in the Lichen-Forming Fungal Genus Oropogon (Parmeliaceae).**
  October 09, 2012
- **Miocene/Pliocene dominated diversification in the lichen-forming fungal genus Melanohalea (Parmeliaceae, Ascomycota) and Pleistocene population expansions**
  October 09, 2012

The current release includes a host of new features and improvements over the previous TreeBASE prototype. New features include:

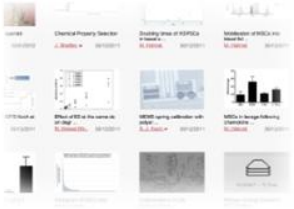
- Richer annotation of metadata (journal DOIs, specimen georeferences, Genbank accession numbers, etc)
- A mapping between taxon labels and taxonomic names in uBio and NCBI for improved normalization of names
- The ability to visualize and edit trees using Phylowidget
- The ability to search on tree topology
- Persistent and resolvable URIs for data objects in TreeBASE (i.e. studies, trees, matrices) serve as both globally unique identification numbers and resource locators. These can be included in articles and on researcher's websites, making access to TreeBASE data only a click away
- Data are delivered in several serializations, including **NEXUS** and **NeXML**
- A special URL gives journal editors and reviewers anonymous advanced access to data
- Programmatic access to the data using the **PhyloWS API**. Queries are expressed in URLs using PhyloWS syntax and can return results in RDF as RSS 1.0 feeds, which means that users can set their favorite RSS Reader to fetch all new TreeBASE studies that satisfy a particular query (e.g. return all studies published in "Systematic Biology," or return all trees that include "Homo sapiens," etc)

As of December 2011, TreeBASE contains 2,946 publications written by 6,106 different authors. These studies analyzed 5,717 matrices and resulted in 8,462 trees with 465,762 taxon labels that mapped to 82,043 distinct taxa.

http://treebase.org

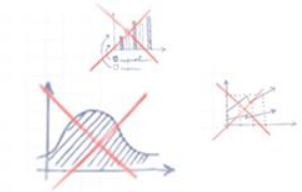
1GB of private space

taggable and easily filtered, your research data is better managed and easy to locate



Unlimited public space

upload to your heart's content
the more - the better



Publish negative data

all published research is citable

Upload all formats

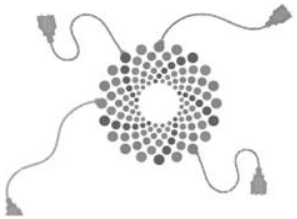
all research outputs welcomed – images, graphs, videos, datasets

Quick & simple upload

upload to your heart's content
the more - the better

Cloud based

Secure and accessible from anywhere



API

An API enables other programs to make use of figshare's functionality and content. Like with all things figshare, the API is open so that it can be accessed and used by anyone.

www.figshare.com

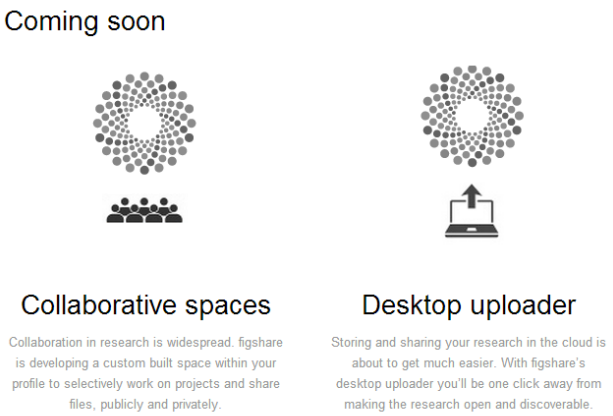
figshare Browse Upload Sign up Login

Webová služba umožňující sdílení dat

Jednoduchá registrace → 1 GB soukromého prostoru

Neomezený prostor pro veřejný obsah

Propojení Figshare s ORCID



Licence ve Figshare

Soukromé úložiště – nevyžaduje licenci

Veřejné úložiště

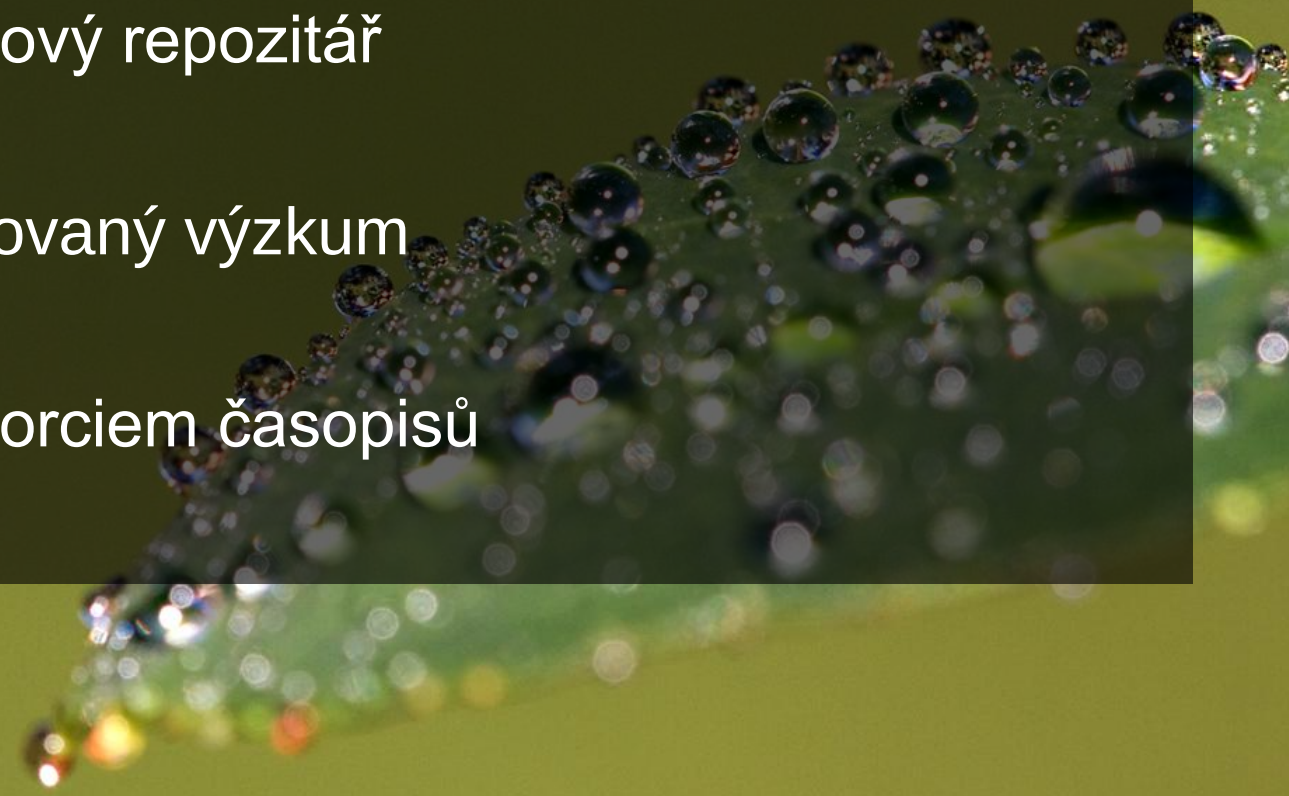
- CC-BY (tabulky, grafy, média, postery, papery, soubory)
- CC0 (datasety) - CC0 is recommended for data and databases and is used by hundreds of organizations. It is especially recommended for scientific data. Although CC0 doesn’t legally require users of the data to cite the source, it does not take away the moral responsibility to give attribution, as is common in scientific research.

http://datadryad.org

Mezinárodní datový repozitář

Základní a aplikovaný výzkum

Spravován konsorciem časopisů



Buzzdata Private Public Use Cases Tools & API Log In Get Started

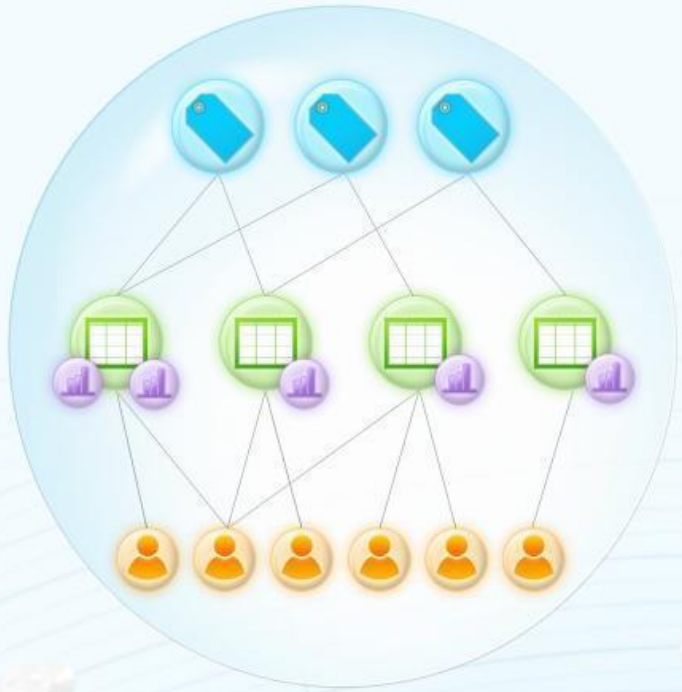
Visual, Smart Executive Summaries.

BuzzData gives you the analysis, visualization & sharing tools you need to find and share the story in your data.

We make it easy to move beyond files to the key facts and figures

Watch the video

BuzzData: Explained



Hive: Your place on the web to share, discover and tell stories with data. Each Hive has a unique name (yourname.buzzdata.com) and contains Datarooms, People and Topics. Hives can be either Private (by invitation only) or Public (open and discoverable via Google searches).

Dataroom: Like folders, only better. Add files, visualizations and links. Tag it with Topics and invite People to check out your work or contribute. See who does what in the activity and comment stream.

Topic: A way to link Datarooms on the same subject together, making it easier to find data. Like a Twitter hashtag, a Topic homepage lets you see all the activity across that Topic.

People: Every person in a Hive has a user profile and dashboard where you can see what's happening with Datarooms, Topics and other People that you're following.

Data: Upload any file type (Excel, CSV, PPT, Word, PDF, KMZ, STATA, etc.) to a Dataroom up to 500MB per file. Developers can get programmatic access to data via our API.

Tools: Coming soon! A range of tools available within Datarooms to help you visualize and work with your data.

Find out what we've learnt about the marketing of data and maximizing return on effort.

Sign up for our weekly newsletter.

There are so many data tools and products out there, it's hard to know which one to use.

www.buzzdata.com

OpenAIREplus

Projekt navazující na OpenAIRE

30 měsíců, 41 partnerů + 3 mezioborové vědecké komunity

Cíle:

1. Propojit publikace s daty a systémy financování
2. Podpořit OA v ERA
3. Vybudovat a provozovat dlouhodobě udržitelné služby

Zdroj: http://dspace.vsb.cz/bitstream/handle/10084/90459/tkacikova-dsugcz2012.pdf

Nástroje na sdílení workflow

Nástroje/platformy pro sdílení surových dat, metodik, postupů, experimentů apod.

MyExperiment - http://www.myexperiment.org

MethodBox - https://www.methodbox.org/

Sysmo-DB - http://www.sysmo-db.org

Enhanced Publications

Obohacené publikace - publ. prolinkované s doplňujícími informacemi / materiály, které se k publ. vztahují.

Př. NARCIS - http://www.narcis.nl

3. listopadu 2012 od 9:00 v NTK

Recycle public sector data to open data


QUESTION MARK

Děkuji za pozornost!

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Technical Library

Tereza Simandlová | @kliste
tereza.simandlova@techlib.cz

Open Access aneb Open your mind! 2012, NTK, 24. 10. 2012

Použité obrázky

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Technical Library

[01] http://www.flickr.com/photos/vixon/116447718/

[02] http://www.flickr.com/photos/neeravbhatt/6995946039/

[03] http://www.flickr.com/photos/31856336@N03/6864002856

[04] http://www.flickr.com/photos/benbeck/3556047352/

[05] http://www.flickr.com/photos/stian_olsen/4411060707/

[06] http://www.flickr.com/photos/feuilllu/5310038972

[07] http://www.flickr.com/photos/guillaumebrialon/3288150201

[08] http://www.flickr.com/photos/sarahreido/3120877348

[09] http://www.flickr.com/photos/juhansonin/5818589649

[10] http://www.flickr.com/photos/getbutterfly/6317955134

[11] http://www.flickr.com/photos/theklan/474695210

[12] Screeshot z Edinburgh DataShare http://datashare.is.ed.ac.uk

[13] Screenshot z TreeBASE http://treebase.org

[14, 15] Screenshot z Figshare http://www.figshare.com

[16] http://www.flickr.com/photos/giuli-o/3421327165

[17] http://www.flickr.com/photos/flod/4870548767 a logo DataDryad.org

[18] Screenshot části obrazovky z Buzzdata

[19] http://www.flickr.com/photos/75166820@N00/26633645 a logo OpenAIREplus http://bit.ly/OAplus

[20] http://www.flickr.com/photos/ian_ruotsala/5218939075

[21] http://www.flickr.com/photos/smartfat/187462753

[22] Logo Big Clean - http://bigclean.techlib.cz/default/files/download/id/3/big-clean-logo.png

[23] http://www.flickr.com/photos/kareneliot/2710464400/

[24] http://www.flickr.com/photos/c04pics/4786851632