9-24-2012

PURR: A Research Data Curation Service Model Using Hubzero

Courtney E. Matthews  
*Purdue University, courtneyearlmatthews@gmail.com*

Michael Witt  
*Purdue University, mwitt@purdue.edu*

Follow this and additional works at: [http://docs.lib.purdue.edu/lib_fspres](http://docs.lib.purdue.edu/lib_fspres)

Part of the [Archival Science Commons](http://docs.lib.purdue.edu/lib_fspres), [Collection Development and Management Commons](http://docs.lib.purdue.edu/lib_fspres), [Scholarly Communication Commons](http://docs.lib.purdue.edu/lib_fspres), and the [Scholarly Publishing Commons](http://docs.lib.purdue.edu/lib_fspres)

Recommended Citation

[http://docs.lib.purdue.edu/lib_fspres/57](http://docs.lib.purdue.edu/lib_fspres/57)

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
PURR: A RESEARCH DATA CURATION SERVICE MODEL USING HUBZERO

Courtney Earl Matthews
Digital Data Repository Specialist
Purdue University
AGENDA

- PURR Background
- Data Management Planning
- PURR Workflow
- Demonstration
- Services / Support
PURR is a hub that provides a research data and collaboration solution for Purdue faculty, graduates, and staff.

Research data - spreadsheets, images, output from sensors and instruments, transcripts, surveys, software source code and tools, video, and observation logs

PURR provides:

• Collaborative Research Space
• Long term Preservation
• Support and Services
• Data Management Plan resources and consultation
DMP REQUIREMENT

GRANT FUNDING AGENCIES

National Science Foundation

National Institutes of Health

National Endowment for the Humanities

NASA Earth Sciences Program

Centre for Disease Control

DMP - https://research.hub.purdue.edu/dmp
NSF DMP Guide - https://research.hub.purdue.edu/overview
PUL = Purdue University Libraries

ITaP = Information Technology at Purdue

OVPR = Office of the Vice President for Research

SPS = Sponsored Program Services

Funded 3-year HUBzero hosting plan (OVPR, Libraries, ITaP)

PURR Steering/Executive Committee:
• Dean of Libraries
• Vice President for Information Technology (CIO)
• Vice President for Research
A **project** is a dedicated working space on PURR for you to collaborate and prepare data for publication and curation for a research project or study.

See [https://research.hub.purdue.edu/kb/AboutPURR/what_is_a_project](https://research.hub.purdue.edu/kb/AboutPURR/what_is_a_project)

A **dataset publication** is a collection of files and metadata you choose from your project to publish.

See [https://research.hub.purdue.edu/publications?](https://research.hub.purdue.edu/publications?)
PURR Workflow mapped to OAIS Reference Model

**SIP - Uncurated Data**  Submission Information Package (SIP) created in HUB

**DIP - Curated Data**  Dissemination Information Package (DIP) created and disseminated on HUB

**AIP - Curated Data**  Archival Information Package (AIP) created and stored in OAIS archive
PROJECT
CREATE PROJECT - 1

[Image of a computer screen showing the Purdue University Research Repository (PURR) interface for a project named 'Networks and Matrix Computations (nmcomp)']

Details:
- Project title: Networks and Matrix Computations (nmcomp)
- General project by David F. Gleich
- File successfully uploaded
- Files:
  - File name: gisich-flickr-2006(...).amat
    - Size: 144.32MB
    - Uploaded: 02/22/2012 9:57 AM
    - By: Michael Witt
    - Revisions: 1
  - File name: Flickr(...).APUsAT
    - Size: 63KB
    - Uploaded: 02/22/2012 9:56 AM
    - By: Michael Witt
    - Revisions: 2
  - File name: README-purr.md
    - Size: 3KB
    - Uploaded: 02/17/2012 11:58 AM
    - By: David F Gleich
    - Revisions: 2
  - File name: README_gisich-flickr(...).pdf
    - Size: 18KB
    - Uploaded: 02/22/2012 10:08 AM
    - By: Michael Witt
    - Revisions: 1
  - File name: README_gisich-flickr(...).txt
    - Size: 3KB
    - Uploaded: 02/22/2012 10:08 AM
    - By: Michael Witt
    - Revisions: 1

[PURR logo]

[Home, MyHUB, Projects, Get Started, Contact Us]

Copyright © 2011 Purdue University. All Rights Reserved. Powered by HUBzero®, a Purdue project.
PROJECT

CREATE PROJECT - 2

Networks and Matrix Computations (nmcomp)

Edit Project

Edit Info

Edit Team

Edit Settings

Edit Info

nmcomp

Title

Networks and Matrix Computations

About

Picture

Upload new image (any existing image will be replaced)

[Choose File] No file chosen

[Upload]

Save changes  Cancel
**Networks and Matrix Computations (nmcomp)**

**Edit Project**
- Edit Info
- Edit Team
- Edit Settings

**Edit Team**
Add new member(s) as:
- manager
- collaborator

<table>
<thead>
<tr>
<th>Individual</th>
<th>Role</th>
<th>Joined</th>
<th>Delete with group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Courtney E. Matthews</td>
<td>manager</td>
<td>23/09/2012</td>
<td></td>
</tr>
<tr>
<td>David F. Gleich</td>
<td>manager</td>
<td>17/02/2012</td>
<td></td>
</tr>
<tr>
<td>Michael Witt</td>
<td>manager</td>
<td>21/02/2012</td>
<td></td>
</tr>
</tbody>
</table>

**Project Owner:**
David F. Gleich
Networks and Matrix Computations (nmcomp)

Edit Project

Edit Info
Edit Team
Edit Settings

Tips
Public projects:
Public projects are listed in the public project directory, with the basic project information visible to users and visitors of the site.

Why provide grant information? Knowing your awarded grant information allows us to allocate the appropriate amount of data storage space to your project. If you omit this information, you will receive default storage space.

Edit Settings

Access
- Project is hidden from search and listings (private project)
- Anyone can find this project in search and listings and view its basic information (public project)

Is this project supported by a research grant?
Fill in some information about your grant. If we can verify your award, we'll automatically increase the storage space available for your project.

Proposal Title:

Principal Investigator:

Funding Agency:

Total Award Budget:

Notify administrator that grant information has been updated

Save changes  Cancel
**Networks and Matrix Computations**

**Manager options:**
- Edit project
- Invite people to join
- Learn about projects

**Publications**

This project has yet no publications. **Start a new publication**

**How the publication process works...**

1. **Choose and arrange your content**
   - Select content from your project files. This may be a single file or multiple files bundled together. You may also add supporting documents e.g. a user guide.

2. **Describe publication and submit for review**
   - Next compose your publication page, adding title, abstract, description, authors and other metadata. You may also add tags and screenshots.

3. **Publish, archive, or save for review**
   - When draft is ready, you may release your work publicly, archive it, or save draft for internal review. Public release comes with a digital object identifier and requires administrator approval.
CREATE PUBLICATION - 2
CREATE PUBLICATION

Networks and Matrix Computations
General project by David P. Gleich

Publications

- Graph of Flickr Photo-Sharing Social Network Crawled in May 2006

Edit Description

- Title: REQUIRED
  - Graph of Flickr Photo-Sharing Social Network Crawled in May 2006

- Synopsis: REQUIRED
  - Graph of the Flickr photo-sharing social network crawled in May 2006 containing a graph with 520,078 nodes and 9,697,214 edges. Crawlers are distributed as a DOT file with NAME file with code to read file in Python and MATLAB.

- Abstract / Description: REQUIRED
  - Flickr is a popular online-community for sharing photos, with millions of users. This graph is representative of its social network, in which the node set $V$ represents users, and the edge set $E$ is such that $(u, v) \in E$ if and only if $u$ and $v$ have edited the same photo in this community. The crawl began with a single user and continued until the total personalized PageRank on this set of uncrawled nodes was less than 0.0001. The result of the crawl was a graph with 520,078 nodes and 9,697,214 edges.
CREATE PUBLICATION - 4
CREATE PUBLICATION - 5

Purdue University Research Repository | PURR

Networks and Matrix Computations (nmcomp)
General project by David F. Gleich

Publications

Add Gallery Images

Publication Gallery:

GleichBклр.APlusAT.gif

Edit
CREATE PUBLICATION - 9

You are here: Projects > Networks and Matrix Computations > Publications > Graph of Flickr Photo-Sharing Social Network...

Networks and Matrix Computations (nmcomp)
General project by David F. Gleich

Publications > dataset "Graph of Flickr Photo-Sharing Social Network Crawled in May 2006"

Version 1.0 (draft)

Title: Graph of Flickr Photo-Sharing Social Network Crawled in May 2006
Version label: 1.0 (default)
Version number: 1
Created: 02/22/2012 (59 minutes ago)
Created by: Michael Witt (mwitt)
Status: draft
Released: 11/30/1999 (1 decade ago)
URL: http://researchhub.purdue.edu/publications/15027?dev

What's Next?
- Publish your work for the world to see! A published version gets a Digital Object Identifier (DOI) and is included in publication listing and search.
- Want to permanently store publication data without public release? You can save it as a "dark archive" with an ARK issued for reference.
- Draft is ready but you want to hold on with publication? Let others in your team know that draft is ready by finalizing and sharing it.
- Changed your mind? You can cancel and delete this draft version.

Click on version label to edit.
Graph of Flickr Photo-Sharing Social Network Crawled in May 2006

By David F Gleich
Purdue University.

Graph of the Flickr photo-sharing social network crawled in May 2006 returning a graph with 926,878 nodes and 9,837,214 edges. Dataset is distributed as a GMAT file with README file with code to read file in Python and MATLAB.

Abstract: Flickr is a popular online-community for sharing photos, with millions of users. This graph is representative of its social network, in which the node set V represents users, and the edge set E is such that (u, v) is in E if and only if a user u has added user v as his/her contact. We start with a crawl extracted from Flickr in May 2006. This crawl began with a single user and continued until the total personalized PageRank on the set of uncrawled nodes was less than 0.0001. The result of the crawl was a graph with 926,878 nodes and 9,837,214 edges.
Networks and Matrix Computations

General project by David F. Gleich

Publications

dataset "Graph of Flickr Photo-Sharing Social Network Crawled in May 2006" Version 1.0 Review

Important notices:

- The publication will be reviewed by the administrator before it goes live.
- A new official DOI (digital object identifier) will be issued with this version release. This means that once this version information is published, it cannot be changed.
- Please carefully review all listed and shared options on this page, and make sure that all your collaborators approve of this publication.
- This publication version is being released with public access; anyone can find the publication page via search and viewing and view/download associated content.

Version 1.0

Graph of Flickr Photo-Sharing Social Network Crawled in May 2006

By David F. Gleich

Purdue University

Graph of the Flickr photo-sharing social network crawled in May 2006 returning a graph with 829,873 nodes and 9,831,214 edges. The dataset is distributed as a SMAT file with a README file with code to read file in Python and MATLAB.

Primary Content - Files

- geon_flickr-2006.smat
  - Added at revision 1 (default)

- README_geon_flickr-2006.pdf
  - Added at revision 1 (default)

- README_geon_flickr-2006.txt
  - Added at revision 1 (default)

Gallery
Graph of Flickr Photo-Sharing Social Network
Crawled in May 2006

By David F Gleich
Purdue University

Crawal of the Flickr photo-sharing social network from May 2006 returning a graph with 820,876 nodes and 9,337,214 edges. Dataset is distributed as a SMAT file with
README file with code to read file in Python and MATLAB.

Abstract: Flickr is a popular online-community for sharing photos, with millions of users. This graph is representative of its
social network, in which the node set V represents users, and the edge set E is such that (u, v) is in E if and
only if a user u has added user v as his/her contact. We start with a crawl extracted from Flickr in May 2006.
This crawl began with a single user and continued until the total personalized PageRank on the set of uncrawled
nodes was less than 0.0001. The result of the crawl was a graph with 820,876 nodes and 9,337,214 edges.
SERVICES & SUPPORT

DMP Overview
DMP Guides and Tutorials
DMP Q & A
  • Storage & Pricing
  • Knowledge Base
  • Contact Us

See https://research.hub.purdue.edu/dmp
Grant Information – SPS
Ticketing - ITap
Data Reference – PUL

See https://research.hub.purdue.edu/dmp
430 Purdue grants include PURR in their DMPs

22 of these grants have been awarded

245 registered users

94 research projects

3 published datasets
PURR: A RESEARCH DATA CURATION SERVICE MODEL USING HUBZERO

MORE INFO?

Courtney Earl Matthews
Digital Data Repository Specialist

Research Department
Stewart Center, Room G50
504 W. State Street
West Lafayette, IN 47907-2058
(765) 496-2770
courtneyearlmatthews@purdue.edu
https://research.hub.purdue.edu