Research Data Integrated & Contextualized in the HKU Scholars Hub

香港大學學術庫：研究數據的整合及脈絡化

David T Palmer
Research Data & Records Development Librarian

Fred Chan
Research & Data Services Librarian
The HKU Scholars Hub 香港大學學術庫

• 2005. IR

• 2009. HKU’s Knowledge Exchange (KE)
  – 3rd mission, after Research & Teaching
  – Engage with the public for mutual benefit

• Libraries’ grant proposal to enlarge the Hub to be,
  – Vehicle for HKU KE
  – Make visible and highlight HKU’s research & skill

• IR ➔ Current Research Information System (CRIS)

• 2005. 機構典藏

• 2009. 香港大學知識交流
  – 研究和教學之外的第三使命
  – 與公眾交流以獲得共同利益

• 圖書館遞交提案，以發展學術庫為：
  – 香港大學知識交流的工具
  – 令香港大學的研究和技能更可見

• 機構典藏 (IR) ➔ 科研信息管理系统 (CRIS)
The HUB: IR ➔ CRIS
學術庫：機構典藏 ➔ 科研信息管理系統

• IR
  – Fulltext publications
  – OA, visibility, preservation

• CRIS
  – Research Objects: publications, researchers, projects, facilities…
  – Visibility
  – Decision support
  – Optimisation of funding process
  – TechTransfer
  – Media resource

• 機構典藏
  – 著作全文
  – 開放獲取、可見性、保存

• 科研信息管理系統
  – 研究物件：著作、研究員、項目、設備…
  – 可見性
  – 協助決策
  – 優化資助過程
  – 技術轉移
  – 媒體資源
HKU research, people and expertise are described and contextualized by more than publications.

Missing in action: Research data

Adapted from: Yiu, ACH. (2016, July). HKU Scholars Hub: From IR to CRIS. Presented at the 1st U21 Librarian Group Summit and University Library Pioneer Conference, Shanghai, China.
Entities are inter-linked

- Facilitate discovery
- Enhance visibility
- Trace the research process

Augmented with external sources

Researchers (Staff & students) are responsible for:
- Manage & document data
- Depositing data

University is responsible for:
- Providing facilities (data repository)
  ➢ Integration with Hub publications, grants, authors
- Training, support, advice

Note: The policy is in English, Chinese translation here is only for reference.
Library responsible for infrastructure
圖書館需要提供基礎設施

• Hardware & software
• UI for deposit
• Indexing & display of metadata
• Training, publicity, support

• 硬件和軟件
• 存放介面
• 索引和顯示元數據
• 培訓、宣傳、支援
We chose the Hub 選擇學術庫作為數據典藏

• Can rapidly create new research objects ➔ research data
• Can easily integrate with other Hub objects ➔ publications, grants, researchers, etc.
• Can easily make deposit page
• HKU already familiar with Hub
• No added costs

• 快捷地建立研究物件 ➔ 研究數據
• 容易跟學術庫其他物件整合 ➔ 著作、研究資助、作者等…
• 容易建立存放頁面
• 大學已經熟悉學術庫
• 沒有額外成本
What is Research Data Management?

HKU now requires RDM. RDM is a general term covering how you organize, structure, store, and care for the information used or generated during a research project. It includes the points below. The benefits of RDM, and the number of funders, journals, and institutions requiring RDM are many and growing.

- Planning how your data will be collected, stored and cared for -- before, during, and after your research project -- many now require data management plans as part of proposals. Many journals as a prerequisite to publishing an require underpinning datasets to be openly deposited. This is a deal with information on a day-to-day basis over the lifetime of a project.
- How you manage your data over the longer term -- what you do with it after your research project concludes.

What Services can HKU Libraries Provide?

The Libraries manage the HKU Scholars Hub which allows deposit of research data, storage, controlled access, and preservation. They offer consultations to individual HKU researchers, group training sessions tailored to your department or faculty, and email and telephone support.
數據集現在是學術庫物件之一
## Dataset Description

**Author**
Liang, Hai
Fu, King-wa

**Contact**
Fu, King-wa

**Date of Dataset Creation**
2015-07-07

**Description**
Replication is an essential requirement for scientific discovery. The current study aims to generalize and replicate 10 propositions made in previous Twitter studies using a representative dataset. Our findings suggest 9 out of 10 propositions could not be replicated due to the variations of data collection, analytic strategies employed, and inconsistent measurements. The replication data include 4 files: 1. EgoAltProfiles (egos are randomly selected Twitter users; alters are the followers and followers of the egos) 2. EgoNetworks (all egos and their following relationships with alters) 3. Exposure (the aggregated data for testing exposure hypothesis) 4. EgoTimelines (all tweets information posted by egos). All datasets have been anonymized.

**Citation**
Liang, Hai, Fu, King-wa (2015). Data from Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science [Data File]. All data are available from the Harvard Dataverse, DOI: http://dx.doi.org/10.7910/DVN/1L1M26

**Subject (RGC Codes)**
- **E2** — Computing Science & Information Technology
  - **2211** — Information Analytics

**Subject (ANZSRC)**
- **20** — LANGUAGE, COMMUNICATION AND CULTURE
  - **2001** — COMMUNICATION AND MEDIA STUDIES

**Keyword**
- Twitter
- Random sample
- Computational social science
編目員進行科目分類，用戶能在探索系統以中英文搜尋

註：ANZSRC的翻譯由國立臺灣大學陳光華教授提供

<table>
<thead>
<tr>
<th>Subject (RGC Codes)</th>
<th>Subject (ANZSRC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2 — Computing Science &amp; Information Technology — 電腦科學及資訊科技</td>
<td>20 — LANGUAGE, COMMUNICATION AND CULTURE — 語言, 傳播與文化</td>
</tr>
<tr>
<td>2211 — Information Analytics — 資訊分析</td>
<td>2001 — COMMUNICATION AND MEDIA STUDIES — 傳播與媒體研究</td>
</tr>
<tr>
<td></td>
<td>200101 — Communication Studies — 傳播研究</td>
</tr>
<tr>
<td></td>
<td>200104 — Media Studies — 媒體研究</td>
</tr>
</tbody>
</table>

DOI for Dataset: http://dx.doi.org/10.7910/DVN/L1MJZ6

Affiliations
1. Univ Hong Kong, Journalism & Media Studies Ctr, Hong Kong, Hong Kong, Peoples R China
數據集連結至作者頁面
Can online opinion reflect public opinion? An investigation into the interplays between online opinion, public opinion, and mass media

<table>
<thead>
<tr>
<th>Project Title</th>
<th>Can online opinion reflect public opinion? An investigation into the interplays between online opinion, public opinion, and mass media</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Investigator</td>
<td>Dr Fu, King Wa (Principal investigator)</td>
</tr>
<tr>
<td>Co-Investigator(s)</td>
<td>Dr Chau Michael Chiul Lung (Co-Investigator)</td>
</tr>
<tr>
<td>Duration</td>
<td>30</td>
</tr>
<tr>
<td>Start Date</td>
<td>2014-07-01</td>
</tr>
<tr>
<td>Completion Date</td>
<td>2016-12-31</td>
</tr>
<tr>
<td>Amount</td>
<td>484250</td>
</tr>
<tr>
<td>Conference Title</td>
<td></td>
</tr>
<tr>
<td>Presentation Title</td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td>online opinion, new media, political participation, public opinion, mass media</td>
</tr>
<tr>
<td>Discipline</td>
<td>Journalism and Media</td>
</tr>
</tbody>
</table>
連結相關著作

The University of Hong Kong

The HKU Scholars Hub 香港大學學術庫

Dataset

Title
Data from: Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science

Liang, Hai 1
Fu, King-wa 1

Date of Dataset Creation
2015-06-07

Description
Replication is an essential requirement for scientific discovery. The current study aims to generalize and replicate 10 propositions made in previous Twitter studies using a representative dataset. Our findings suggest 6 out of 10 propositions could not be replicated due to the variations of data collection, analytic strategies employed and inconsistent measurements. The replication data include 4 files: 1. EgoAltarProfiles (egos are randomly selected Twitter users; altars are the followers and followees of the egos) 2. EgoNetworks (all egos and their following relationships with alters) 3. Exposure (the aggregated data for testing exposure hypothesis) 4. EgoTimelines (all tweets information posted by egos) All datasets have been anonymized.

Citation
Liang, Hai, Fu, King-wa. (2015). Data from: Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science [Data File]. All data are available from the Harvard Dataverse, DOI: http://dx.doi.org/10.7910/DVN/L1MJJZB

Subject (RGC Codes)
E2 — Computing Science & Information Technology — 電腦科學及資訊科技
L 2211 — Information Analytics — 資訊分析
Replication is an essential requirement for scientific discovery. The current study aims to generalize and replicate 10 propositions made in previous Twitter studies using a representative dataset. Our findings suggest 6 out of 10 propositions could not be replicated due to the variations of data collection, analytic strategies employed, and inconsistent measurements. The study's contributions are twofold. First, it systematically summarized and assessed some important claims in the field, which can inform future studies. Second, it proposed a feasible approach to generating a random sample of Twitter users and its associated ego networks, which might serve as a solution for answering social-scientific questions at the individual level without accessing the complete data archive.
Replication is an essential requirement for scientific discovery. The current study aims to generalize and replicate 10 propositions made in previous Twitter studies using a representative dataset. Our findings suggest 6 out of 10 propositions could not be replicated due to the variations of data collection, analytic strategies employed, and inconsistent measures. The study’s contributions are twofold: First, it systematically summarized and assessed some important claims in the field, which can inform future studies. Second, it proposed a feasible approach to generating a random sample of Twitter users and its associated ego networks, which might serve as a solution for answering social-scientific questions at the individual level without accessing the complete data archive.
## Dataset

<table>
<thead>
<tr>
<th>Title</th>
<th>Data from: Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science</th>
</tr>
</thead>
</table>
| Author| Liang, Hai ¹  
Fu, King-wa ² |
| Contact| Fu, King-wa |
| Date of Dataset Creation | 2015-05-07 |
| Description | Replication is an essential requirement for scientific discovery. The current study aims to generalize and replicate 10 propositions made in previous Twitter studies using a representative dataset. Our findings suggest 6 out of 10 propositions could not be replicated due to the variations of data collection, analytic strategies employed, and inconsistent measurements. The replication data include 4 files: 1. EgoProfiles (ego profiles are randomly selected Twitter users, alters are the followers and followers of the ego); 2. EgoNetworks (all egos and their following relationships with alters); 3. Exposure (the aggregated data for testing exposure hypothesis); 4. EgoTimelines (all tweets information posted by egos). All datasets have been anonymized. |

### Citations

Google Scholar: 2

Liang, Hai, Fu, King-wa. (2015). Data from: Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science [Data File]. All data are available from the Harvard Dataverse, DOI: http://dx.doi.org/10.7910/DVN/L1MJZ6

#### Subject (RGC Codes)

- E2 — Computing Science & Information Technology — 電腦科學及資訊科技
  - 2211 — Information Analytics — 資料分析

#### Subject (ANZSRC)

- 20 — LANGUAGE, COMMUNICATION AND CULTURE — 語言·傳播與文化
  - 2001 — COMMUNICATION AND MEDIA STUDIES — 傳播與媒體研究
    - 200104 — Media Studies — 媒體研究

#### Keyword

- Twitter
- Random sample
- Computational social science
在Google Scholar的引用次数为2次
從Harvard Dataverse的DOI獲取數據集
從Web of Science搜索香港大學研究者其PLOS著作相關的數據集加入學術庫
• Organization-Enhanced
• Publication Name
• Year Published
組織的記錄顯示其研究者累積的數據集
數據存放頁面，將在2016年9月正式運作
Integrated & Contextualized 整合及脈絡化

• Data, part of the research cycle
  – PI, department, grant, publication, citations ➔ impact!

• Discovery
  – Discovery on one, can lead to discovery on all
  – More chance for all to be read & cited
  – Many access points

• Funders & admin can understand better the impact of their sponsoring $$

• 數據是研究過程的一環
  – 首席研究員、部門、研究資助、著作、引用 ➔ 影響力！

• 探索
  – 由個別物件引伸至所有相關資訊
  – 更大機會被閱讀和引用
  – 多個存取點

• 資助機構及行政人員能更了解研究資金的影響力
### Extending DSpace ➔ DSpace-CRIS

- **Relational Tables**: top level objects & attributes

<table>
<thead>
<tr>
<th></th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Publications</strong></td>
<td>- Article title</td>
</tr>
<tr>
<td></td>
<td>- Publication year, etc.</td>
</tr>
<tr>
<td><strong>Researchers</strong></td>
<td>- Name(s)</td>
</tr>
<tr>
<td></td>
<td>- Department, etc.</td>
</tr>
<tr>
<td><strong>Datasets</strong></td>
<td>- Research project</td>
</tr>
<tr>
<td></td>
<td>- Citations, etc.</td>
</tr>
</tbody>
</table>

- **UI**: Mash-up to show integrated display
- **Developed with our partners**, CINECA of Italy
- **Open Source**: [https://wiki.duraspace.org/display/DSPACECRIS/DSpace-CRIS+Home](https://wiki.duraspace.org/display/DSPACECRIS/DSpace-CRIS+Home)
擴展 DSpace ➔ DSpace-CRIS

• 相關表格：頂級物件及屬性

著作
• 文章標題
• 出版年份等

研究員
• 姓名
• 部門等

數據集
• 研究計劃
• 引用次數等

• 混搭以顯示整合介面
• 與意大利的CINECA合作開發

• 開源軟體：https://wiki.duraspace.org/display/DSPACECRIS/DSpace-CRIS+Home
References 參考書目


Acknowledgement

• The authors would like to thank the HKU Scholars Hub team for their dedicated work and concerted effort in improving the Hub.

• This PPT & PDF are licensed under the Creative Commons Attribution “CC-BY” license. This license permits use of this work, so long as attribution is given. For more information about the license, visit http://creativecommons.org/licenses/by/3.0/

• Archived at http://hub.hku.hk/handle/10722/230618

David T Palmer
http://hub.hku.hk/cris/rp/rp00001
http://orcid.org/0000-0001-5616-4635

Fred Chan
http://hub.hku.hk/cris/rp/rp01995
http://orcid.org/0000-0002-3832-8357