10-2014

Research Data Management and Curation Aspirations at NTU and SMU Libraries

Wei Yeow CHENG
Nanyang Technological University, Singapore

Tint Hla Hla HTOO
Singapore Management University, thhhtoo@smu.edu.sg

Citation

Cheng, Wei Yeow and Goh Su Nee. 2014. "Research Data Management and Curation Aspirations at NTU and SMU Libraries." Presentation at Libraries for Tomorrow Conference, Singapore, October 14.

LAS conference 2014 - Research Data Management and Curation Aspirations at NTU and SMU Libraries

- Cheng Wei Yeow and Goh Su Nee, Nanyang Technological University Libraries
- Tint Hla Hla Htoo, Singapore Management University Library

LAS conference 2014-
NTU Libraries Research Data Management

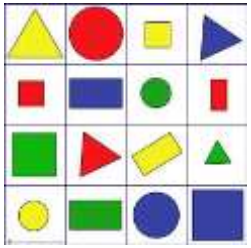
Cheng Wei Yeow and Goh Su Nee
(Scholarly Communication Group)

Discover NTU Libraries

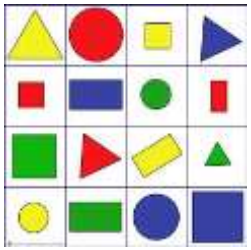
Outline

NTU Libraries Research Data Management

Strategy → **H**indrance → **A**ction → **P**lan → **E**xcellent **S**ervice



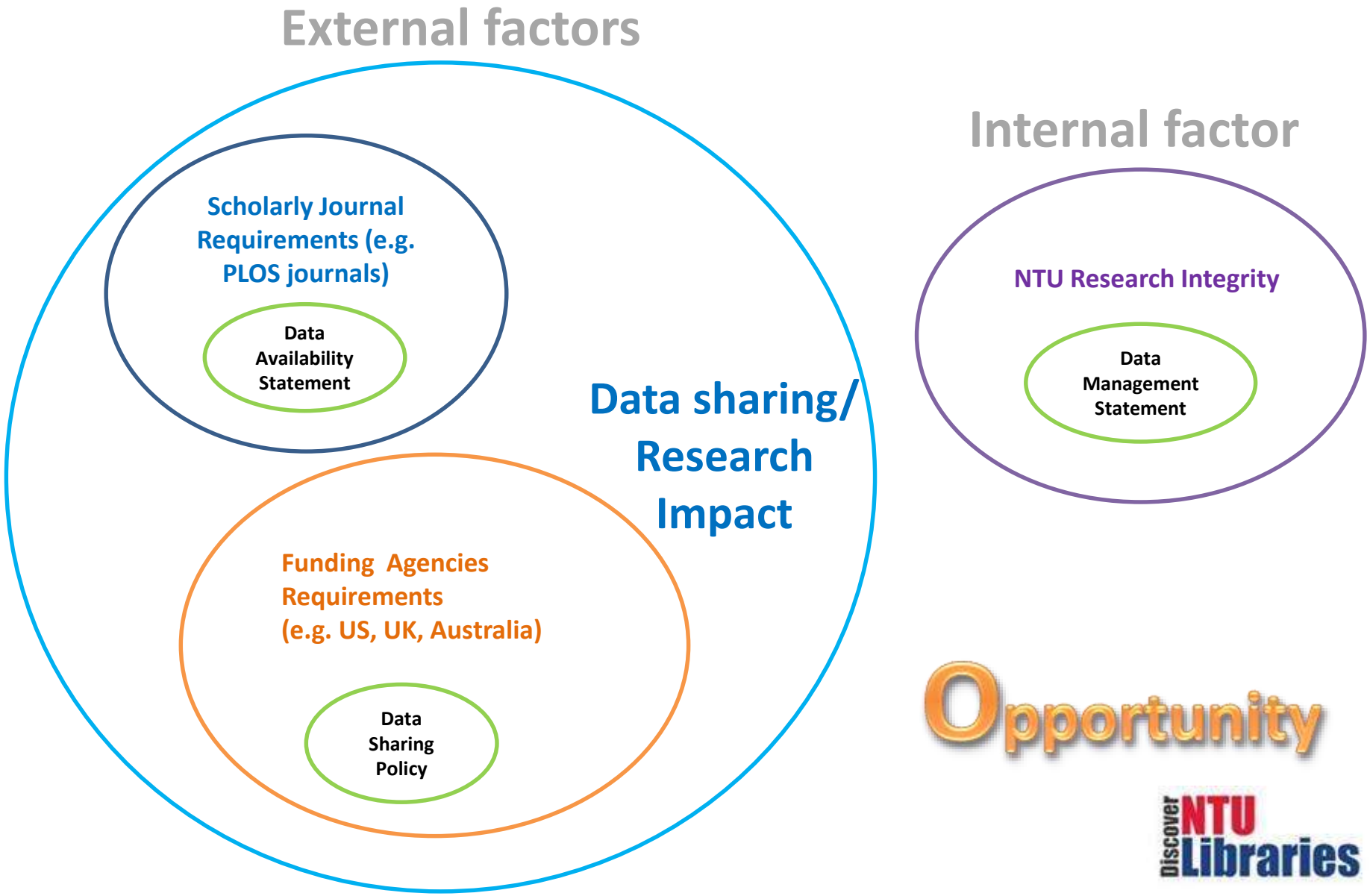
SHAPES



Discover NTU Libraries

Strategy

The shape of things to come



Opportunity

Discover NTU Libraries

Strategy

Shape up to shape a new service

Lib re-org for Technical Services Group in 2014

Expanded the Scholarly Communication Group in 2014

New team- Research Data Management (April 2014)

2 former technical services librarians

Staff 1 (0.5 FTE) + Staff 2 (<0.5 FTE)
[almost everyone is a dual role librarian: operational role and subject role]

Discover NTU Libraries

Hindrance

in many shapes

Challenges

No buy-in from stakeholders

No institutional requirement

No local funding agency requirement

Lack of systematic capture of research data

No relevant experience

Tight manpower situation

No research fund to develop Data Management Service

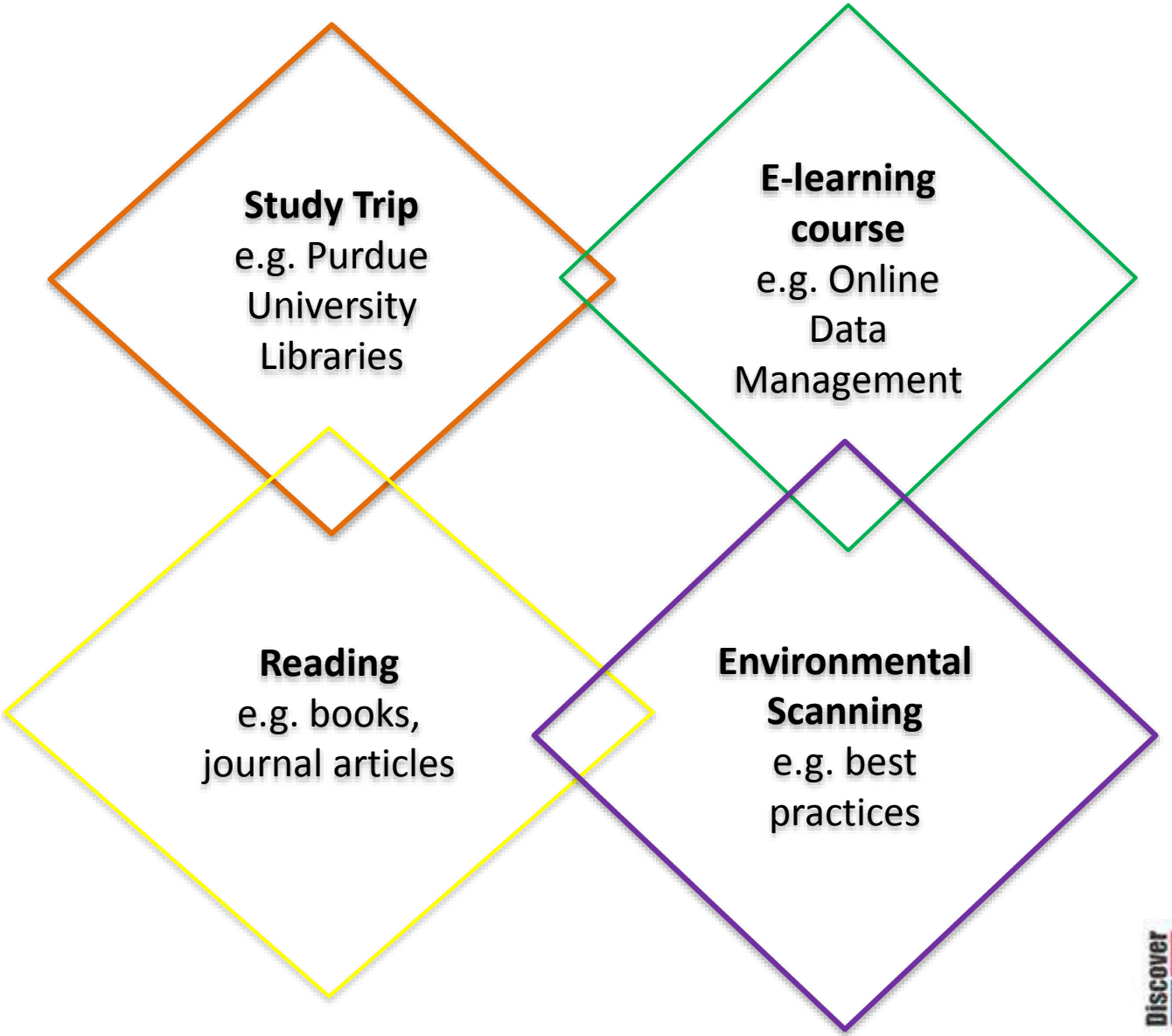
Limitations of infrastructure

Source: http://allaboutbaby2011.blogspot.sg/

Discover NTU Libraries

Action

in the shape of



Discover NTU Libraries

Plan

Knocking Research Data Management in NTU into shape

Roadmap

2014
1. **Policy**
2. **Infrastructure**
3. **Education**

2015
1. Policy
2. Infrastructure
3. Education

2016
1. Policy
2. Infrastructure
3. Education

Activities:

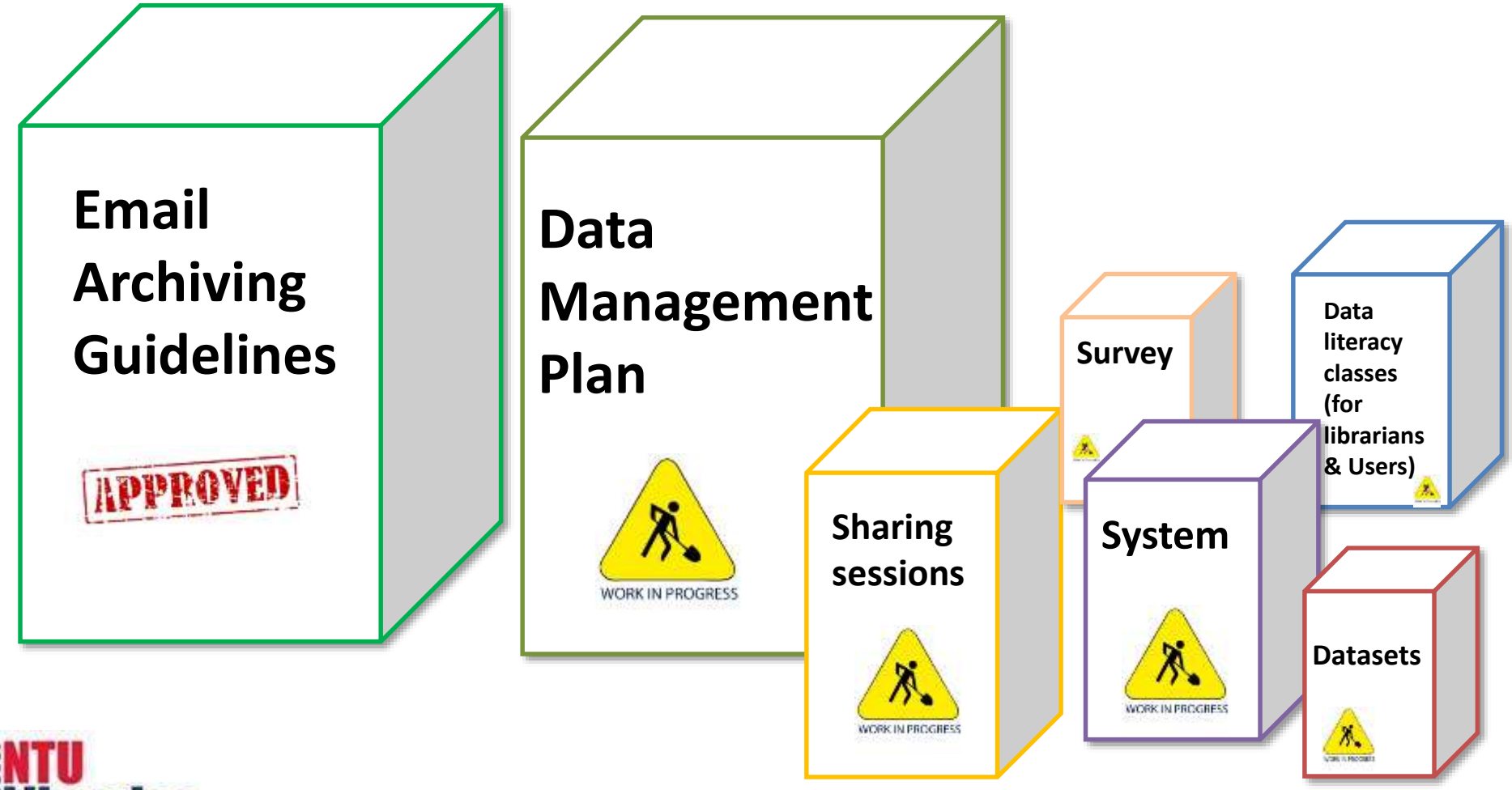
- **Email Archiving**
- **Data Management Plan**
- **Data literacy classes**
- **Survey**
- **Sharing sessions**

Discover NTU Libraries

Plan

Beginning to take shape…

and we are in great shape, so far

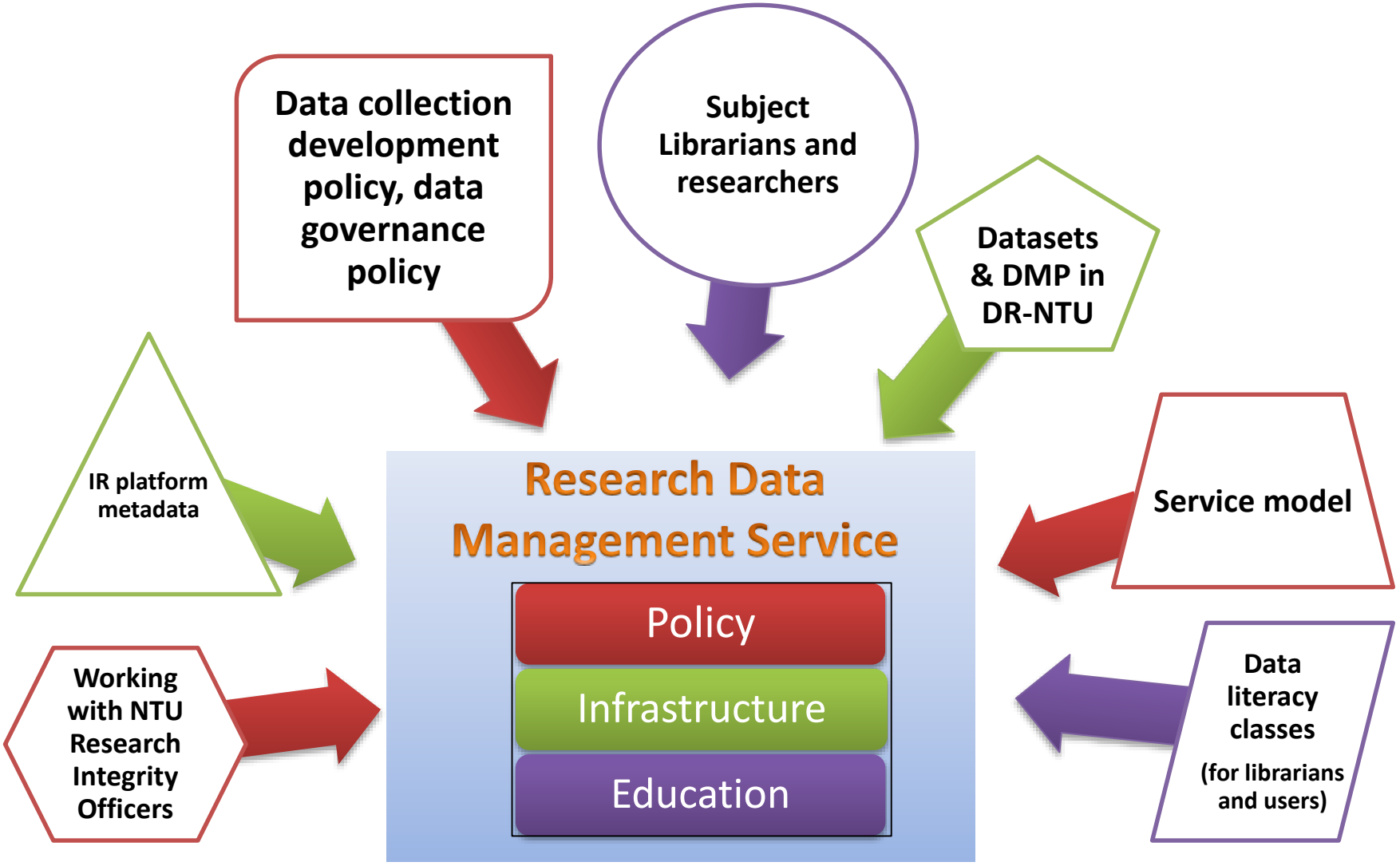


Discover NTU Libraries

Plan

Discover NTU Libraries

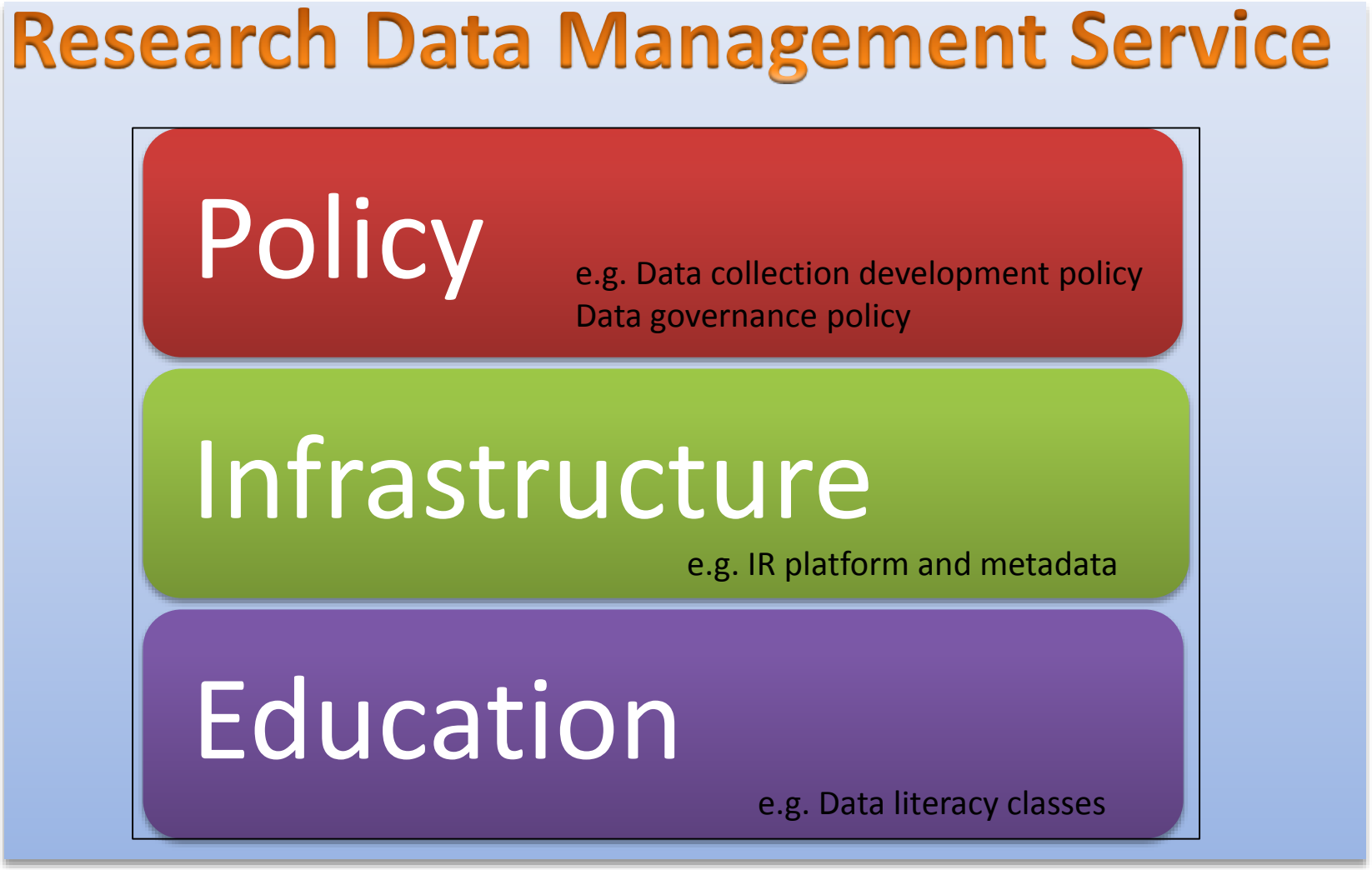
Shaping in progress



Excellent Service

Discover NTU Libraries

Aiming to be in good shape



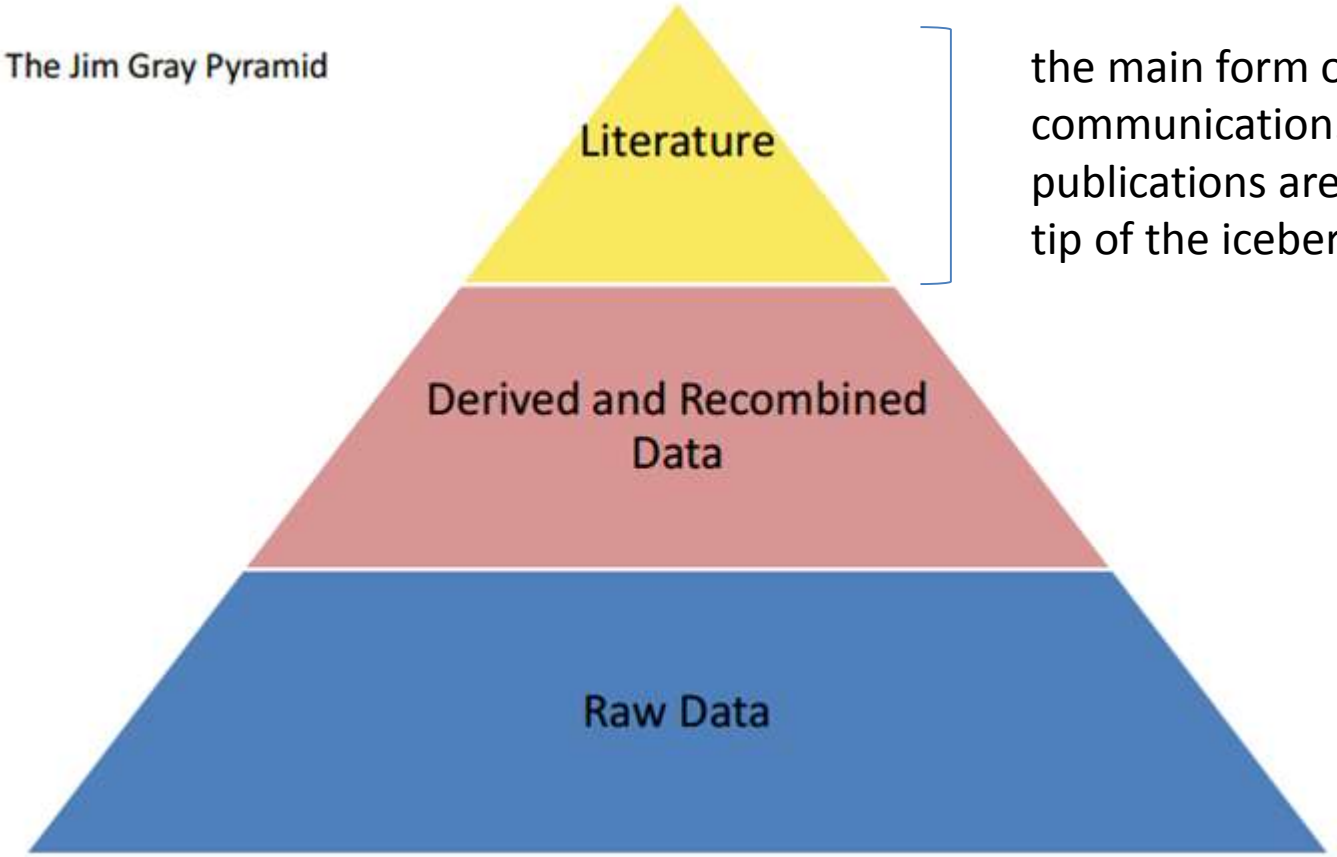
LAS conference 2014
Data Curation – SMU Libraries

Tint Hla Hla Htoo
Research Data Services Librarian

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Why are libraries doing data curation?

SMU
SINGAPORE MANAGEMENT UNIVERSITY



the main form of scholarly communication. But publications are only the tip of the iceberg

Research data are often:

1. Unavailable,
2. Unfindable, if available at all
3. Uninterpretable. If available AND findable at all
4. Not re-usable, if available, findable, interpretable at all

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Availability and reusability of research data have big social and economic impact on our society.

Science will progress faster.

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Data Availability & Reuse - The Drivers

Research Funders - *adopt data sharing policy or mandate*

Publishers - *Data availability policy or mandate requires authors to make data and material available to readers, as a condition of publication.*

Researchers – *(With varying degrees) Share data to increase impact and visibility of research, promotes innovation and potential new data uses, etc.*

Data Archives & Libraries *– support infrastructure and related services*

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Research Data in School of Information Systems (SMU)

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Research Data – School of Information Systems (SMU)

Jing Jiang

Assistant Professor

School of Information Systems
Singapore Management University
80 Stamford Road
Singapore 178902

Phone: (+65) 6828 0785
Email: jingjiang at smu dot edu dot sg

[Home] [Research] [Publications] [Code & Data] [Group]

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Code & Data

Code

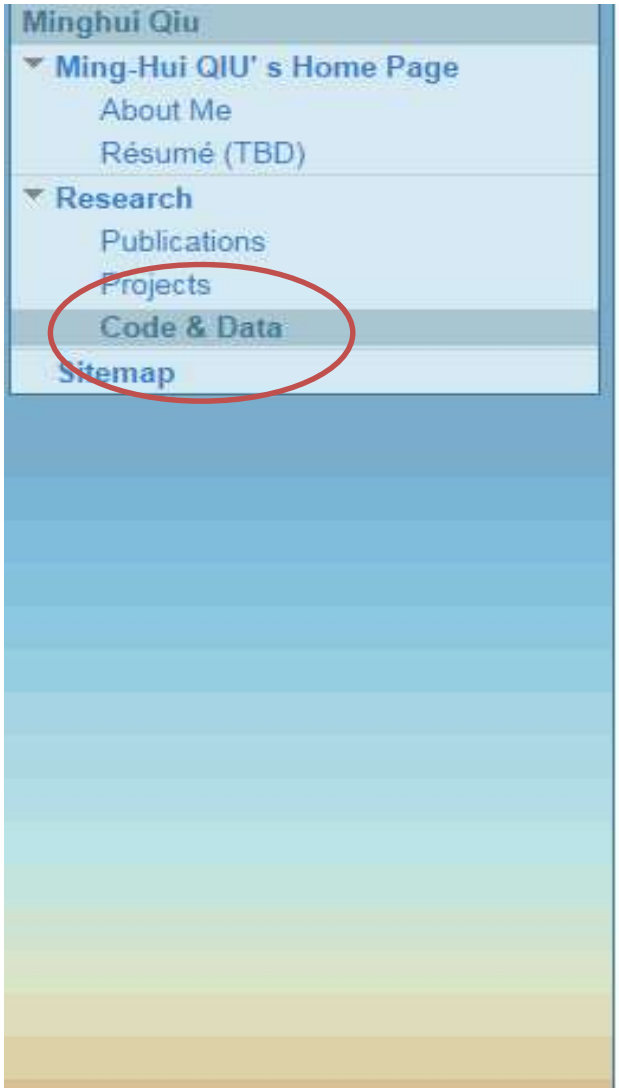
- BioTokenizer.pl: As a first step to many information retrieval and natural language processing tasks, tokenization is the process of seperating text into individual tokens that each convey some semantic meaning. For English, in most cases, tokens are equivalent to words. For biomedical text, there are often names and symbols of various types of biomedical entities, such as genes, proteins, chemicals, etc. The special characters contained in these names and symbols make it harder to identify meaningful tokens than in normal English text. This piece of code in Perl implements a number of tokenization heuristics we have studied in the following paper:
    - Jing Jiang and ChengXiang Zhai. An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval*, 10(4-5):341-363, October 2007.
- Domain Adaptive Logistic Regression: I have implemented a number of domain adaptation techniques that I explored in my PhD thesis in this toolkit.

Data

Code & Data by My Students

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Research Data – School of Information Systems (SMU)



Code & Data

- Debatepedia dataset Download
  - Reference: Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, a Debatepedia. EMNLP'13.
- Topic Expertise Model
  - Code: Java Code (Github TEM)
  - This package implements Gibbs sampling for Topic Expertise Model for j
  - Reference: CQARank: Jointly Model Topics and Expertise in Community
- PMF Model for Mining User Relations
  - Code: Code (Github)
  - 6 data sets from CreateDebate Download
  - Reference: Mining User Relations from Online Discussions using S
- B-LDA (Joint Behavior-Topic Model)
  - Code: Java Code (Github B-LDA)
  - We propose an LDA-based behavior-topic model (B-LDA) which jointly m the model on on-line social network settings such as microblogs like Twitt them are rich.
  - Reference: It's Not What We Say But How We Say Them: LDA-based Be Austin, Texas, USA, May, 2013.
- Twitter-LDA
  - Code: Java Code (Github Twitter-LDA)
  - The original setting in Latent Dirichlet Allocation (LDA), where each word single tweet is more likely to talk about one topic. Hence, Twitter-LDA (T- Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan and Xiaoming Li. Cor of the 33rd European Conference on Information Retrieval (ECIR'11) " to where it captures background words in tweets.

Research Data – School of Information Systems (SMU)

Publications

Scalable Code Clone Detection

Our studies and others' have noticed that on average more than 20% of code in large programs is cloned c
maintenance cost and subtle software defects. The goal of our research is to scalably and accurately detec
evolutions and migrations among large programs, and manage them properly to facilitate program underst
applications, such as code refactoring, bug detection, and plagiarism detection, can stem from code clone de

- DECKARD: A Code Clone and Clone-Related Bug Detection Tool
  - Checkout the latest versions at `git://github.com/skyhover/Deckard.git`
  - The git repository can also be viewed at https://github.com/skyhover/Deckard
  - Or download a stable source package in the 7z format: version 1.2.3
  - Or try out a parallelized version that makes Deckard run faster on a multi-core machine: versio
- *Understanding the Genetic Makeup of Linux Device Drivers*, by Peter Senna TSCHUDIN, Laurent David LO, Julia LAWALL, and Gilles MULLER. In the proceedings of the 7th Workshop on Program Systems (PLOS '13), Farmington, Pennsylvania, USA, 2013. [on ACM DL, pdf]
- *Active Refinement of Clone Anomaly Reports*, by Lucia, David LO, Lingxiao JIANG, and Aditya Budi. ternational Conference on Software Engineering (ICSE '12), Zurich, Switzerland, 2012. [on IEEE Xplo
- *Automatic Mining of Functionally Equivalent Code Fragments via Random Testing*, by Lingxiao JIAN ceedings of the 18th International Conference on Software Testing and Analysis (*ISSTA '09*), Chicag ACM DL, on ACM DL, pdf, slides.pdf]

LIBOL

A Library for Online Learning Algorithms

Download

Version 0.3.0

MATLAB/Octave Interfaces, core functions in matlab and C/C++.

File Name	Version Number	Release Date
Source Code	0.3.0 Beta (.zip) [901KB]	12 Dec 2013
Manual	LIBOL_manual (.PDF) [4KB]	12 Dec 2013
technical report	LIBOL_TR (.pdf) [118KB]	12 Dec 2013
Datasets	libol_DB1 (.zip) [89MB]	27 July 2013
Datasets	libol_DB1 (.zip) [34MB]	27 July 2013

OLD Versions

Version 0.2.3 (released on 23 Sep 2013)
MATLAB/Octave Interfaces, core functions in C/C++.

File Name	Version Number	Release Date
Source Code	0.2.3 Beta (.zip) [855KB]	23 Sep 2013
Manual	LIBOL_manual (.PDF) [4KB]	23 Sep 2013
technical report	PDF (.pdf) [118KB]	23 Sep 2013
Datasets	libol_DB1 (.zip) [89MB]	27 July 2013
Datasets	libol_DB1 (.zip) [34MB]	27 July 2013

Version 0.2.0 (released on 27 July 2013)
MATLAB Interface, C/C++ implememtatnion for core functions.

Navigation

OLPS

On-Line Portfolio Selection via Machine Learning

Home · Software

Software and Code

- CWMR --- Confidence Weighted Mean Reversion Strategy
  [Project Webpage] [CODE]
- PAMR --- Passive Aggressive Mean Reversion Strategy
  [Project Webpage] [CODE]

The software of our On-Line Portfolio Selection toolbox will be released soon.

Navigation

SBFA

Search Based Face Annotation

WLF: A database of Weakly Labeld Faces on the Web

WLF - Weakly Labeled Faces on the web, is a large-scale real web facial images database, which consists of a total of 714,454 facial images and 6025 persons collected from the internet. There are about 118 images per person on average. The minimal number of facial images per person is 28, and the maximal number is 187.

Click here to download the official WLF database.

Navigation

- About
- WLF Database
- Demo
- Software
- People
- Publication
- Contact

Data Curation in SMU Libraries

- Provide infrastructure and related services to ensure long term availability and access to data
- Institutional Repository to collect data and other research outputs, in addition to publications

Why deposit in IR

- Institutional Archive
- Robust infrastructure
- Compliant with international standards and protocols for maximum discoverability
- Granular access control
- Manage by professionals
- Measure impact by Altmetrics

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Challenges

- Content and Other Policies
- Organization & Description
- Copyright & Licensing
- Limitation with current IR infrastructure (e.g. no permanent identifier issued which is often required for tracking and data citation to demonstrate impact)
- Demonstrating Impact

SMU
SINGAPORE MANAGEMENT UNIVERSITY