FOUNDATIONS of DATA CURATION

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales

School of Information Sciences • University of Illinois at Urbana-Champaign

Workflow and Provenance

(anything profound, and the cool slides, is from Bertram Ludäscher. Everything else is from Renear

V3. Workflow systems

Workflow systems: requirements, examples

Provenance in workflow (& provenance standards)

(cool slides immediately or mediately from (Bertram Ludäscher)

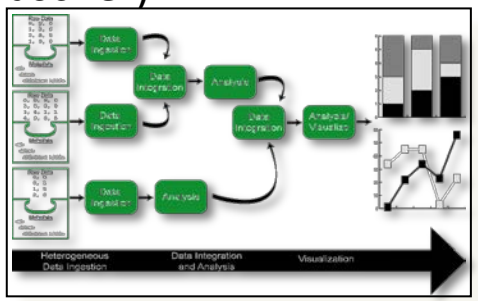
Essential Functions of a Scientific Workflow System (Bertram Ludäscher)

1. **Automate** programs and services scientists already use.
2. **Schedule** invocations of programs and services correctly and efficiently – **in parallel** where possible.
3. **Manage dataflow** to, from, and between programs and services.
4. **Enable scientists** (not just developers) to **author** or **modify** workflows easily.
5. **Predict** what a workflow will do when executed: ***prospective* provenance**.
6. **Record** what happened during workflow execution: ***retrospective* provenance**.
7. **Reveal and query provenance** – how workflow products were derived from inputs via programs and services.
8. **Organize** intermediate and final **data products** as desired by users.
9. Enable scientists to **version**, **share** and **publish** their workflows.
10. **Empower scientists** who wish to **automate additional programs and services themselves.**

These functions (not just dataflow & actors) distinguish ***scientific workflow automation*** from general (scientific) software development.

Src: Tim McPhillips

Scientific Workflows: ASAP (Bertram Ludäscher)



- **Automation**
  - wfs to **automate** computational aspects of science

- **Scaling** (exploit and optimize ***machine*** *cycles*)
  - wfs should make use of **parallel compute resources**
  - wfs should be able handle **large data**

Kepler

swift

- **Abstraction, Evolution, Reuse** *(**human** cycles)*
  - wfs should be easy to **(re-)use, evolve, share**

Taverna

ASKALON

VisTrails

TRIANA

- **Provenance**
  - wfs should capture **processing history**, **data lineage**

➔ traceable data- and wf-evolution

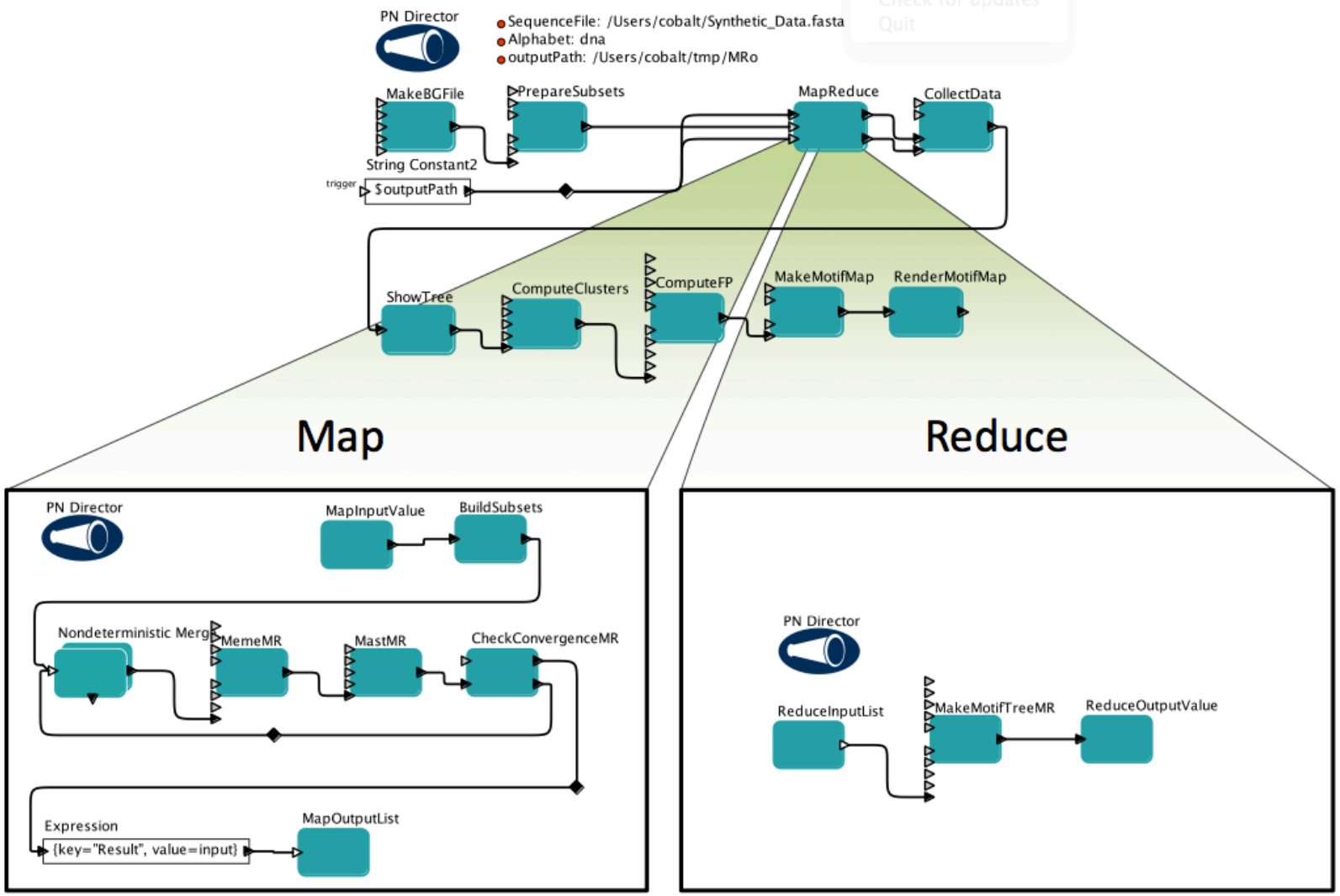
➔ **Reproducible Science**

pegasus

Trident Workbench

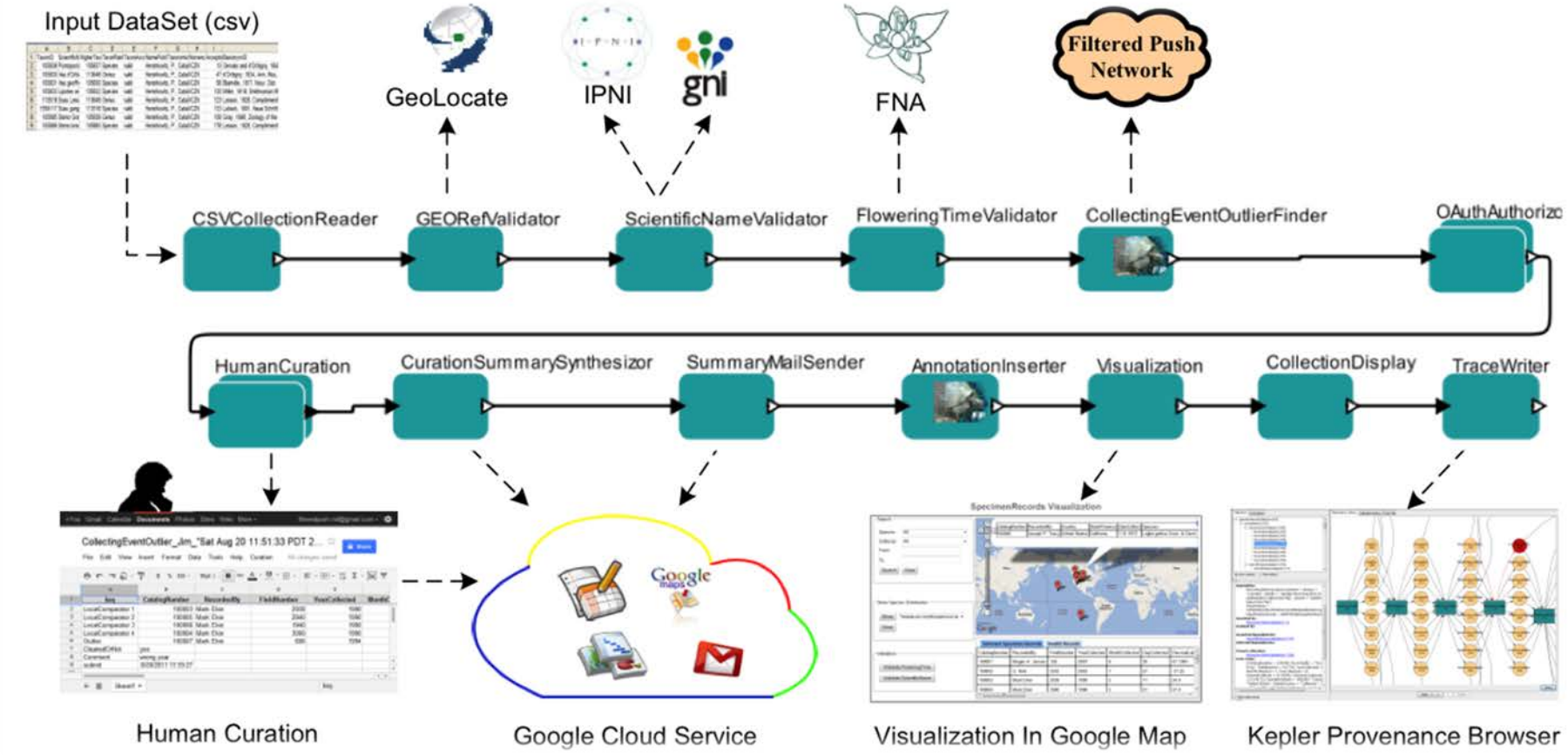
Es war einmal …

Motif-Catcher workflow, implemented in Kepler (Bertram Ludäscher)



S Köhler et al. Improved Motif Detection in Large Sequence Sets with Random Sampling in a Kepler workflow, *ICCS-WS*, 2012

Data Curation Workflows (here: using Kepler)



Museum of Comparative Zoology
Harvard University

UCDAVIS genome center

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

(Bertram Ludäscher)

Yes, scripts are (can be) workflows too!

Reproducible academic publications

This section contains academic papers that have been published in the peer-reviewed literature or pre-print sites such as the ArXiv that include one or more notebooks that enable (even if only partially) readers to reproduce the results of the publication. If you include a publication here, please link to the journal article as well as providing the nbviewer notebook link (and any other relevant resources associated with the paper).

- Automatic segmentation of odor maps in the mouse olfactory bulb using regularized non-negative matrix factorization, by J. Soelter et al. (Neuroimage 2014, Open Access). The notebook allows to reproduce most figures from the paper and provides a deeper look at the data. The full code repository is also available.
- Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss, by A. Gross et al. (Nature Genetics 2014). The full collection of notebooks to replicate the results.
- powerlaw: a Python package for analysis of heavy-tailed distributions, by J. Alstott et al.. Notebook of examples in manuscript, ArXiv link and project repository.
- Collaborative cloud-enabled tools allow rapid, reproducible biological insights, by B. Ragan-Kelley et al.. The main notebook, the full collection of related notebooks and the companion site with the Amazon AMI information for reproducing the full paper.
- A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data, by C.T. Brown et al.. Full notebook, ArXiv link and project repository.
- The kinematics of the Local Group in a cosmological context by J.E. Forero-Romero et al.. The Full notebook and also all the data in a github repo.
- Warming Ocean Threatens Sea Life, an article in Scientific American backed by a notebook for its main plot. By Roberto de Almeida from MarinExplore.

- AtomPy: An Open Atomic Data Curation Environm
Applications, by C. Mendoza, J. Boswell, D. Ajok

Data-driven journalism

- The Need for Openness in Data Journalism, by B
- St. Louis County Segregation Analysis , analysis
Area Is Even More Segregated Than You Probab
Singer-Vine.

```
[3]: display(i)
```

IP[y]: IPython
Interactive Computing

```
In [3]: from IPython.display import SVG
        SVG(filename='python-logo.svg')

Out[3]:
```

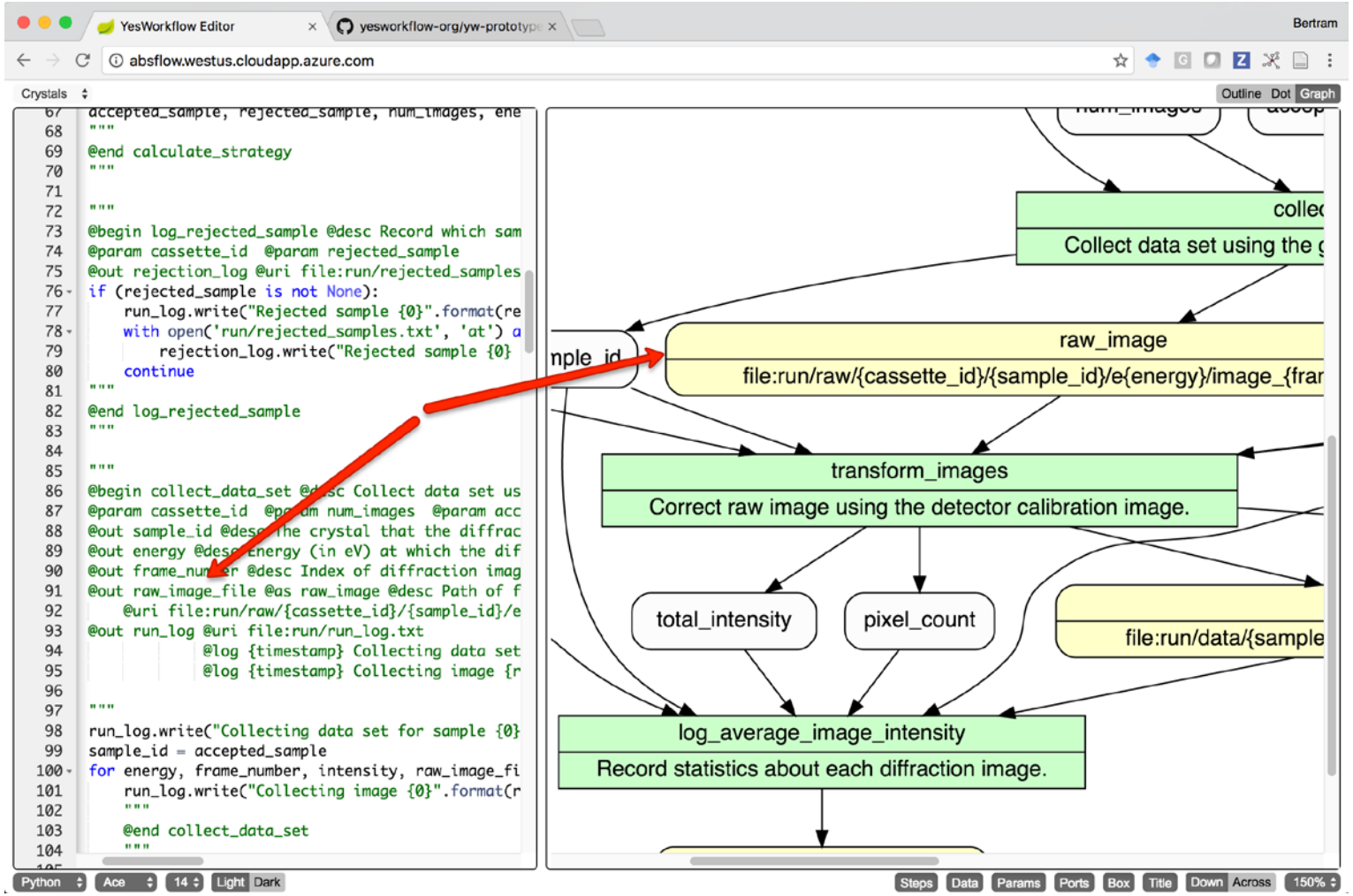
python™

What programming/statistics languages you used for an analytics / data mining / data science work in 2013? [713 votes total]

% users in 2013 | % users in 2012 | % users in 2011

	% users in 2013	% users in 2012	% users in 2011
R (434 voters in 2013)	60.9%	52.5%	45.1%
Python (277)	38.8%	36.1%	24.6%
SQL (261)	36.6%	32.1%	32.3%
SAS (148)	20.8%	19.7%	21.2%
Java (118)	16.5%	21.2%	24.4%
MATLAB (89)	12.5%	13.1%	14.6%
High-level data mining suite (80)	11.2%	not asked in 2012	
Unix shell/awk/sed (79)	11.1%	14.7%	
C/C++ (66)	9.3%	14.3%	
Pig Latin/Hive/other	8.0%		

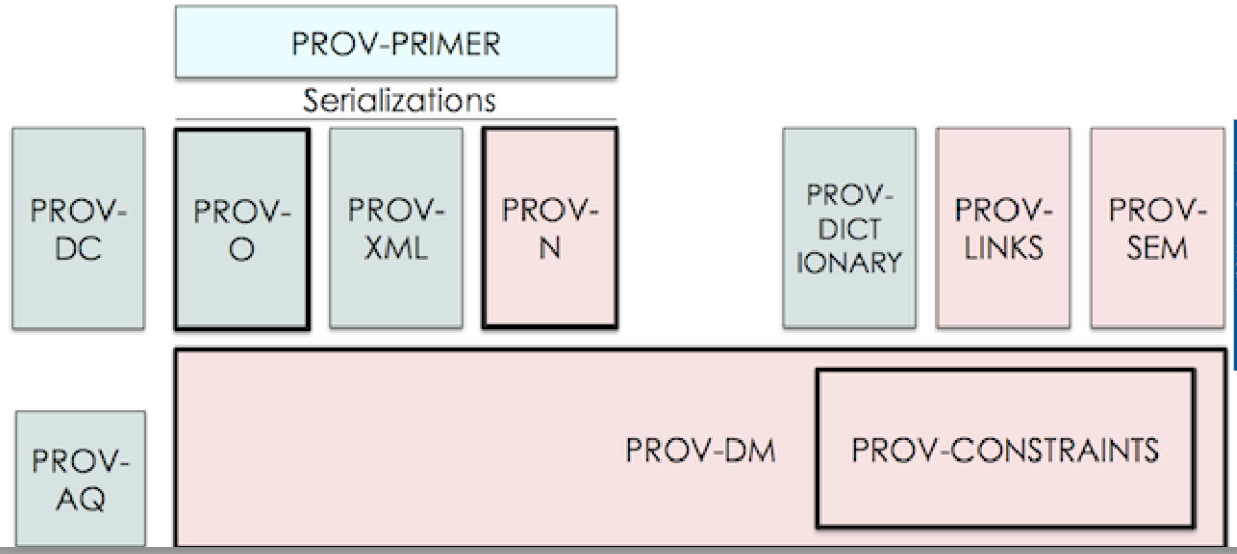
try.yesworkflow.org

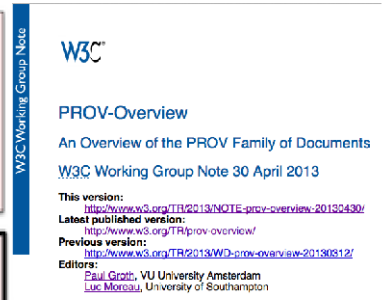


W3C PROV Family of Specifications: Modeling

(Bertram Ludäscher)

- W3C Recommendations
  - PROV Data Model (PROV-DM)
  - PROV Ontology (PROV-O)
  - PROV-Constraints
  - PROV Notations (PROV-N)
- PROV Working Group Notes (selected)
  - PROV-Access and Querying (AQ)
  - PROV Dictionary
  - PROV XML
  - PROV and Dublin Core Mappings (PROV-DC)
  - PROV Semantics (using first-order logic) (PROV-SEM)




W3C Working Group Note

W3C

PROV-Overview

An Overview of the PROV Family of Documents

W3C Working Group Note 30 April 2013

This version:
http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
Latest published version:
http://www.w3.org/TR/prov-overview/
Previous version:
http://www.w3.org/TR/2013/WD-prov-overview-20130312/
Editors:
Paul Groth, VU University Amsterdam
Luc Moreau, University of Southampton

Provenance Analysis and RDF Query Processing, Satya S. Sahoo, Praveen Rao, ISWC, October, 2015.

At least remember this

don’t just sit there typing at the command line,

write a script, and document it

[or, more vividly. . .]

Automate like you are going to live forever

Document like you are going to die tomorrow

— *Michael Sperberg-McQueen*

FOUNDATIONS OF DATA CURATION (IS531)

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales
School of Information Sciences
University of Illinois at Urbana-Champaign

Includes material adapted from work by Carole Palmer, Melissa Cragin, David Dubin, Karen Wickett, Bertram Ludaescher, Ruth Duerr and Simone Sacchi.

Comments and corrections to: renear@illinois.edu.