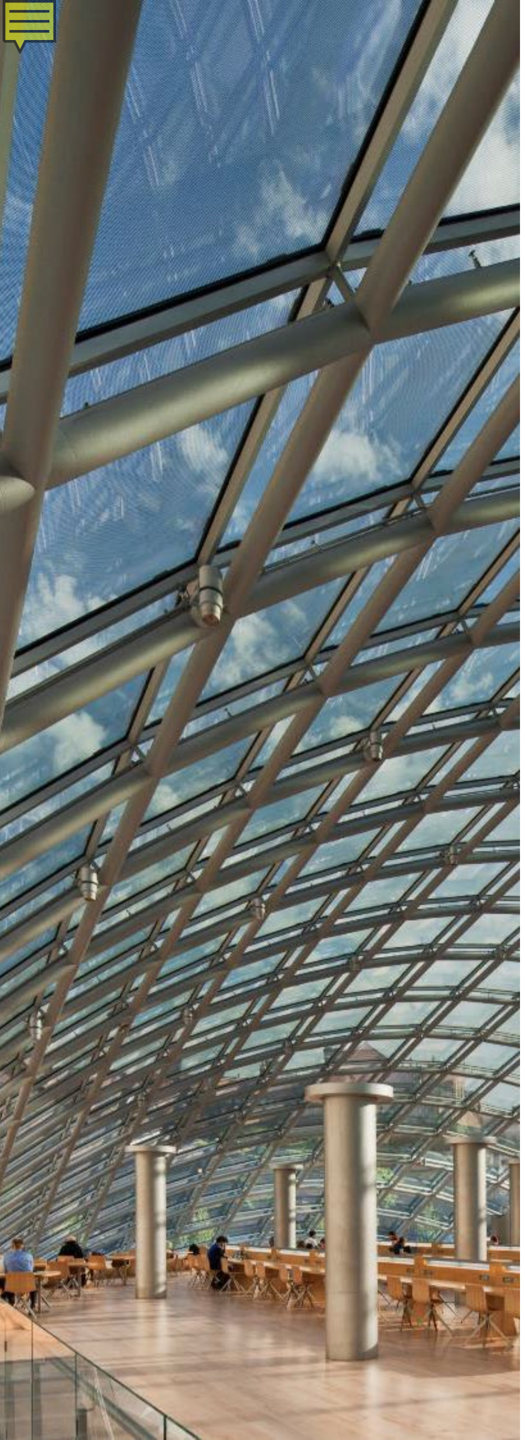


kwalk

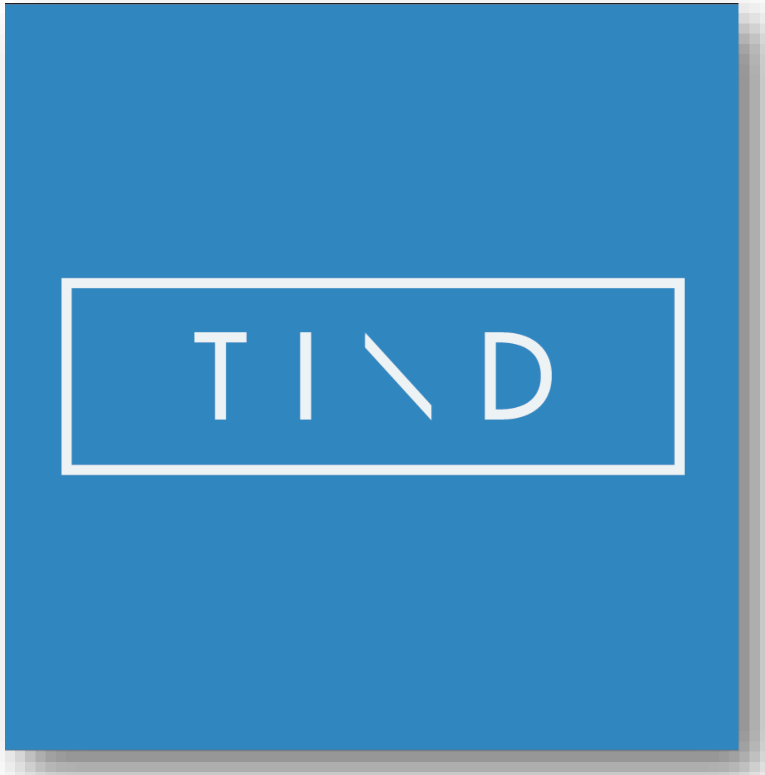
a simple program to crosswalk metadata for repository uploads

Kirsten Vallee

May 14, 2024

UChicago Library

Knowledge@UChicago



- **Repository launched in 2015**
- **Migrated from DSpaceDirect to TIND**
- **About TIND**
  - Official CERN spin-off
  - Built on CERN technology
  - Metadata is structured and stored as MARC
- The k in kwalk = knowledge (not Kirsten)
  - And knowledge is power

Kwalk is built on refertool

created by Keith Waclena, Applications Systems Analyst/Programmer-Lead in the Digital Library Development Center (DLDC) at the University of Chicago Library

a multi-tool for manipulating refer databases

Scenario where kwalk is useful

You need to upload 1,000 items from a source like Lens.org or PLOS journals

You obtain informal metadata either on your own or from another person via the following ways:

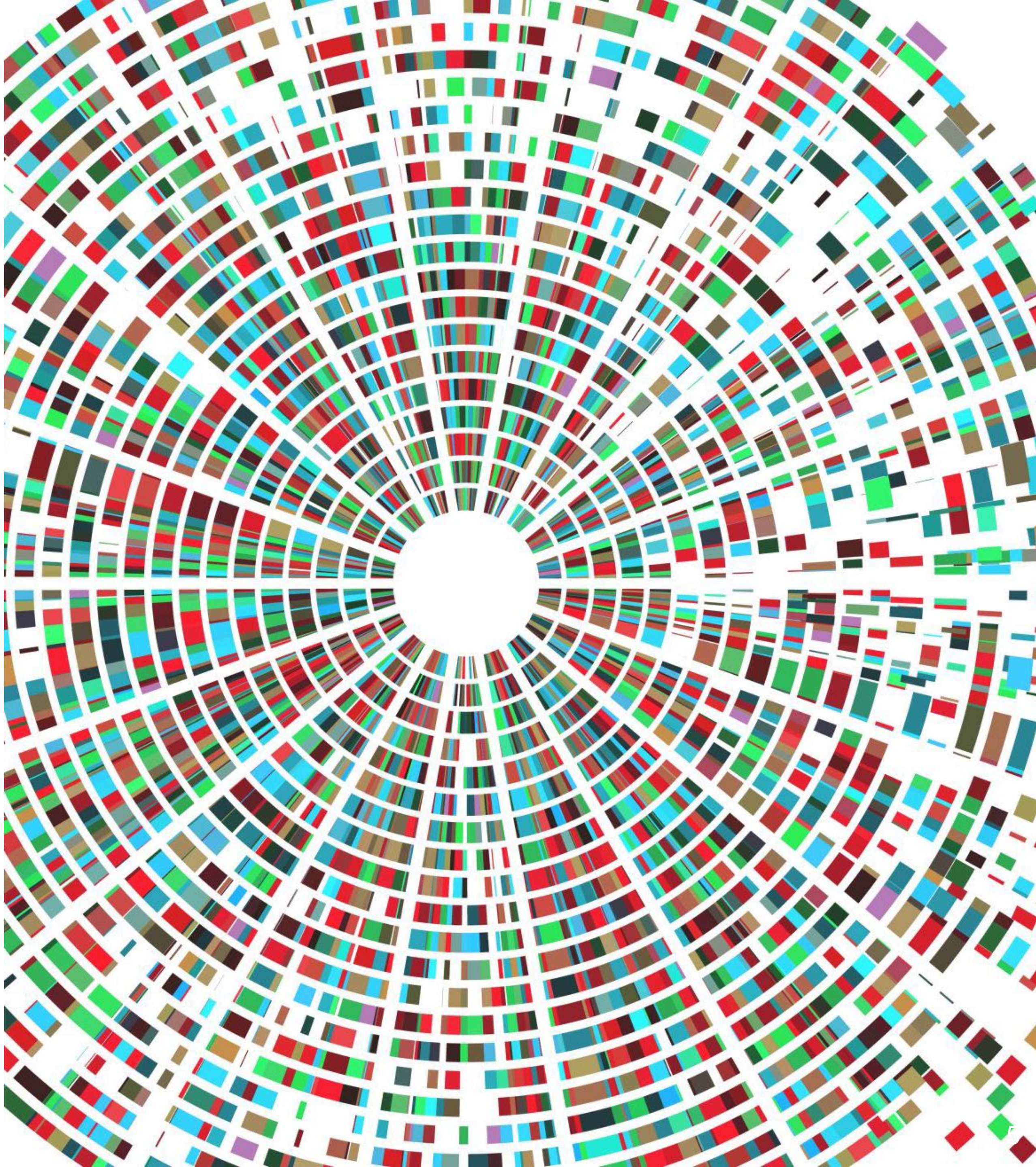
Creating the spreadsheet from scratch

Exporting the data

Web scraping

What can you do with kwalk?

- **Apply a crosswalk to a batch of metadata, for example:**
  - Spreadsheet data > MARC
  - MARC > Dublin Core
  - Spreadsheet data > Dublin Core
  - The possibilities are (maybe) endless
- **Mix and match multiple crosswalks on multiple projects**
- **Transform data**



Possible data transformations – you might need to

rename	Rename all the fields from the invented field names to your repository platform’s field names
add	Add some fields that are missing
exclude	Exclude some fields you don’t want
combine	Combine several fields into one field
modify	Modify the values of date formats or author names in a programmatic way
generate	Generate syntactically correct upload URLs from a simple filename field

Some useful special functions in kwalk

CONVERTDATE

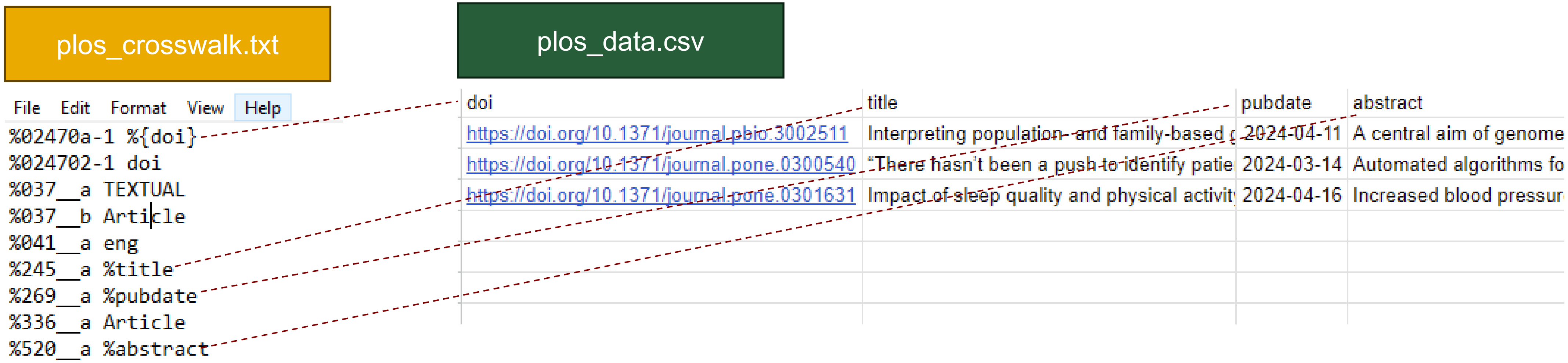
- The `CONVERTDATE` function is used to change the format of a date in some input field
- It takes three arguments
  - The input field name
  - A format that matches the date in that field
  - A format to which we want to convert
    - i.e. `%260__c CONVERTDATE %date %m/%e/%Y %Y-%m-%d`
      - This will convert 05/03/2024 in an incoming date field to 2024-05-03

Combining literal and field name text

- i.e. `%269__a Spring %year`
- i.e. `%FFA__a https://www.lib.uchicago.edu/%thesis_file_name`

How to kwalk .csv to MARC

1. A **crosswalk** in .txt format (i.e. plos_crosswalk.txt)
2. A **database** in .csv format (i.e. plos_data.csv)



plos_crosswalk.txt

```
%02470a-1 %{doi}
%024702-1 doi
%037__a TEXTUAL
%037__b Article
%041__a eng
%245__a %title
%269__a %pubdate
%336__a Article
%520__a %abstract
```

plos_data.csv

doi	title	pubdate	abstract
https://doi.org/10.1371/journal.pbio.3002511	Interpreting population- and family-based g	2024-04-11	A central aim of genome
https://doi.org/10.1371/journal.pone.0300540	"There hasn't been a push to identify patie	2024-03-14	Automated algorithms fo
https://doi.org/10.1371/journal.pone.0301631	Impact of sleep quality and physical activit	2024-04-16	Increased blood pressur

How to kwalk .csv to MARC

01

Open a task management program like Windows PowerShell

02

Navigate to the directory where your .txt and .csv files are located

03

Run `kwalk-tind -c filename.txt dataname.csv > desiredoutputname.csv`

Outcome

02470a-1	024702-1	037__a	037__b	041__a	245__a	269__a	336__a	520__a
https://doi.org/10.1371/journal.pbio.3002511	doi	TEXTUAL	Article	eng	Interpreting population- and family-based genome-wide association studies in the presence of confounding	2024-04-11	Article	A central aim of genome-wide association studies (GWASs) is to estimate direct genetic effects: the causal effects on an individual's phenotype of the alleles that they carry. However, estimates of direct effects can be subject to genetic and environmental confounding and can also absorb the "indirect" genetic effects of . . .
https://doi.org/10.1371/journal.pone.0300540	doi	TEXTUAL	Article	eng	"There hasn't been a push to identify patients in the emergency department"—Staff perspectives on automated identification of candidates for pre-exposure prophylaxis (PrEP): A qualitative study	2024-03-14	Article	Automated algorithms for identifying potential pre-exposure prophylaxis (PrEP) candidates are effective among men, yet often fail to detect cisgender women (hereafter referred to as "women") who would most benefit from PrEP. The emergency department (ED) is an opportune setting for implementing automated identification of PrEP candidates, but there are logistical and practical challenges at the individual, provider, and system level. In this study, we aimed to understand existing processes for identifying PrEP candidates and to explore determinants for incorporating automated identification . . .
https://doi.org/10.1371/journal.pone.0301631	doi	TEXTUAL	Article	eng	Impact of sleep quality and physical activity on blood pressure variability	2024-04-16	Article	Increased blood pressure variability (BPV) is linked to cardiovascular disease and mortality, yet few modifiable BPV risk factors are known. We aimed to assess the relationship between sleep quality and activity level on longitudinal BPV in a . . .

Yes, you can convert from .csv to Dublin Core

The crosswalk would look something like this:

```
%xmlns:dc http://purl.org/dc/elements/1.1/

%xmlns:xsi http://www.w3.org/2001/XMLSchema-instance

%xsi:schemaLocation http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd

%dc:identifier %doi

%dc:language %language

%dc:creator %author1

%dc:creator %author2

%dc:title %title

%dc:identifier %data_url

%dc:date CONVERTDATE SEP=; %pubdate;%B %e, %Y;%Y-%m-%d

%dc:type Text
```

Yes, you can convert from .csv to Dublin Core

The crosswalk would look something like this:

```
%xmlns:dc http://purl.org/dc/elements/1.1/

%xmlns:xsi http://www.w3.org/2001/XMLSchema-instance

%xsi:schemaLocation http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd

%dc:identifier %doi

%dc:language %language

%dc:creator %author1

%dc:creator %author2

%dc:title %title

%dc:identifier %data_url

%dc:date CONVERTDATE SEP=; %pubdate;%B %e, %Y;%Y-%m-%d

%dc:type Text
```

The output would be:

```
%xmlns:dc http://purl.org/dc/elements/1.1/

%xmlns:xsi http://www.w3.org/2001/XMLSchema-instance

%xsi:schemaLocation http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd

%dc:identifier doi:https://doi.org/10.1021/acsnano.3c12581

%dc:language eng

%dc:creator Caillas, Augustin

%dc:creator Guyot-Sionnest, Philippe

%dc:title Uncooled High Detectivity Mid-Infrared Photoconductor Using HgTe Quantum Dots and Nanoantennas

%dc:identifier https://knowledge.uchicago.edu/record/11512/files/Uncooled-High-Detectivity-Mid-Infrared-Photoconductor-Using-HgTe-Quantum-Dots-and-Nanoantennas.pdf

%dc:date 2024-04-04

%dc:type Text
```

Yes, you can convert from .csv to Dublin Core

The crosswalk would look something like this:

```
%xmlns:dc http://purl.org/dc/elements/1.1/
%xmlns:xsi http:/
instance
%xsi:schemaLocati
http://www.openar
http://www.openar
%dc:identifier %d
%dc:language %lan
%dc:creator %auth
%dc:creator %auth
%dc:title %title
%dc:identifier %d
%dc:date CONVERTD
%m-%d
%dc:type Text
```

The output would be:

```
%xmlns:dc http://purl.org/dc/elements/1.1/
/XMLSchema-
enarchives.org/OA
ives.org/OAI/2.0/
g/10.1021/acsnano
ppe
ty Mid-
e Quantum
uchicago.edu/reco
ctivity-Mid-
e-Quantum-Dots-
%dc:date 2024-04-04
%dc:type Text
```

Turn it into an XML file by running:

```
refertool tocsv -h fake-input.db | refertool fromcsv -h | ~/src/kwalk/bin/main -f fake-xwalk.db | refertool toxml -x -r dc:dc -R collection -A xsi:schemaLocation -A xmlns:xsi -A xmlns:dc fake-output.db
```

Questions? Ideas for other uses?

Kirsten Vallee

vallee@uchicago.edu