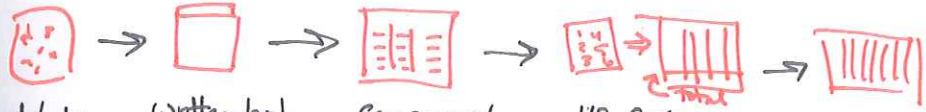
IASSIST
2015

June 3-5, 2015

Steve Ruggles

The History of Data: Technological Change and the Census 1790-2020

DATA CAPTURE



inventory (beer!) — property — people ~~property~~

23 million cases in 1850 to tabulate!

SEETON DEVICE

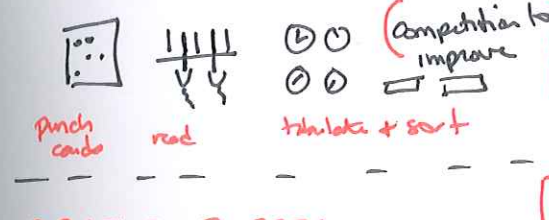
HOLLERITH TABULATOR → IBM

US CENSUS MACHINE → Remington Rand

bought machines from IBM in 1950

1950's used computers! UNIVAC

Univac bought by Remington Rand



CENSUS BUREAU

directly or indirectly spawned two biggest computer companies of early 20th century

- 1960 mail out census form

photo ~~sensing~~ reading for collection

* punchcards

- 1967 invention of GIS
- 1970 mailed out AND mailed back

1/1000 1/10000 microdata for universities !

used different #2 pencils

Compared w/ 1960's data due to compatability

- 1980
- 1990

} faster computers ...

- 2000 data capture outsourced to Lockhead Martin

lots of scanning, lots of errors

- 2010 "was a collasol failure"

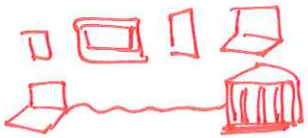
handheld devices ?

ended up w/ Lockhead again due to failures by other contractors, Harris

CANCELED internet response option

↑ CANADA did this no problem

- 2020 ? Not looking good ...



bring your own device

internet response option

neither looking good

US CENSUS TIMELINE

Recover / compare old data

1960-1990

Missing chicago data from 1960

data discovered in a cave in Kentucky (really)

Build logitudinal study from 1940's on...

} CLIP

link to other existing microdata

truly an extraordinary moment to be a data nerd !

CENSUS BUREAU

- drove huge innovations in computing in 20TH century
- charged recently with outsourcing
- still ahead in data (depth, breadth, free)

SESSION A5 - RDM SERVICES

Jungwon Yang
University of Michigan

SERVICES @ UMich
- teach each other
- data ed working group
- program-oriented RDS (not project

"Stakeholder Analysis for RDM services for public policy researchers"

TENURE → faculty's main interest → stakeholders → { publish, grants

84 faculty
47 research
37 clinial

Journals
- dairy & lotsa other data
  - stats
  - micro / macro
  - codebooks

Grants
- applied for / 48 have federal grants (NSF + NIH)
- IRB big component
  - sensitive data

<CLOSUP>
- survey group on policy
- had non data curation specialist!

ISSUES
- reliable / supported repository
- preparing data for sharing
- anonymizing sensitive data

FUTURE PLANS
- guide for anonymizing sensitive data
- data education for students
- work on data repository

WANT TO SHARE DATA (pressure from APSA)

Maya Ishida
U. Manitoba

Sarah Williams
UIUC

Developing Research Data Services Vision(s):
An analysis of ARL libraries

Lots of libraries developing services

① Driver docs (fed ag's + memos)

② ARL libraries web pages + strategic doc

③ Semi-structured interviews w/ admins

72% had RDS main web page
48% other related pages

PRELIMINARY FINDINGS

ALL SERVICES
- 24% DMP assistance
- Consults / instruction
- Data deposit / repos
- Storage
- Sharing / re-use
- Best practices / info dissemination
- Archiving / preservation
- Data processing / Analysis
- 3% Metadata

<Strategic planning>
- 56% mention data in library-wide strategic plan
- 7% have data-focused vision

INTERVIEWS

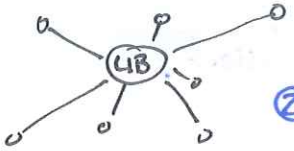
COLLABORATION AND ENGAGEMENT
- library can't be only place supporting data
- partner not just service

ENABLING RESEARCH ACROSS LIFECYCLE

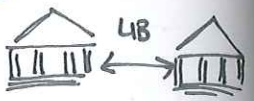
STRONG EXPERTISE
- integrate and bring others on board

VISION THEMES

① library should be HUB of RDM network



VISION SUGGESTIONS



② provide support across disciplines

U Minnesota | Amy West, Amy Neeser |

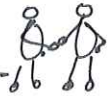
A Coordinated, Decentralized Approach to RDM services: from education to everyday

we don't have a choice but to do things as cooperatively as we can

UMinn
- varying levels of expertise
- culture of decentralization
- dept/working group silos

WHAT WORKS

Never an end to relationship building



- at least one person who facilitates
- Communication { share content!
- freedom to experiment

< maintain communication & workflows between groups >

. . . . . .

Use this to tailor teaching by discipline

Quantitative
Qualitative
Simulation
Other

rough breakdown of data on campus

DISCIPLINARY VARIATIONS

* Acknowledge diversity
* Focus on users
* Develop general content customize as needed
* Iterative simultaneous efforts to train
  - library staff
  - researchers

< STAFF TRAINING >

just start!

like an internship training

Slightly different training than researchers

RESEARCHER TRAINING

← leverage RCR requirement

734 attendees since 2009!

① wave
come one come all!

② targeting different audiences
students, faculty
bring other librarians on board

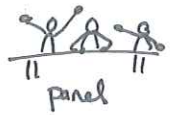
SPECIALIZED
- ☐ Civil engineering course ↓repurposed
- ☐ life science online
- ☐ social science workshops
- ☐ grad students
- ☐ DMP consults

It takes time ...

... and you're never done

SESSION B5 - Building on Common Ground

Lizzy Rolando - Georgia Tech
Kelly Chatain - University of Michigan
Bethany Anderson - UIUC



Kelly

Survey ~~Group~~ (Research Center) hired to do data archiving/records management

↑
20-30 studys per year
~140 employees

Records Management

- ☐ Assess the situation
  - regulatory/compliance issues
- ☐ Records Survey
- ☐ Research/Business Task Analysis
- ☐ Strengths and Challenges

→ ~~Retain~~ Retention Schedule
- granular
- accessible
- annually reviewed
- actionable

engage researchers upstream in data lifecycle

⇒ Targetted Guidelines
- *Email
- *File naming/digitization
- *Archive checklist
- *Project closeout
- Trainings
- Annual updates

OUTREACH
- Records Day
- "From The Archives" articles

{ encouraging good behaviors
improve documentation
and more positive changes! }

(SESSION B5)

Lizzy R.

it started with a Thesis...

that had a dataset...

this was a mess

ARCHIVES would maintain in current form
would want to fix in DATA SERVICES

how do we handle records? how is it different b/t different services?

limited technical resources → Borrow processing from ARCHIVES
Learn about digital sustainability from RDS

Can we build 1 system for Archives + Data?

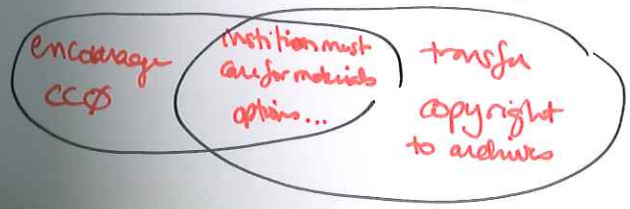
THEORY

DATA
+ prioritize use over preservation
+ context valuable in respect to original material

ARCHIVES
+ prioritize preservation over use
+ Context valuable on its own

< different priorities >

need to determine retention and deaccessioning



messy stuff here...

"don't make process too easy or it will be dumping ground"

Borrow workflows from archives

ingest → preservation → access

Trying to share tools as well...

ARCHIVES ♥'s DATA SERVICES

<Bethany> Appraising and Preserving Data in Context

<Appraisal means identifying and selecting>

Many reasons to appraise

storage is going to be a problem w/out careful selection

~~Expensive~~ expensive to keep/maintain everything

How to appraise

- mission alignment
- institutional compliance
- scientific/historical value
- uniqueness
- potential for redistribution
- non-replicability
- economic factors
- contextual documentation and metadata

Haas, Samuels, Trippel Simmons
Appraising the Records of Modern Science and Technology
1985

[What context is needed for data?]

methods, how data was acquired, etc.

PRESERVE MEMORY ← both from data

C3: Data Sharing Behavior and Policy

"The road to data sharing is paved w/ good intention"

Laurence Horton [London School of Economics & Political Science] & Astrid Recker [Leibniz Institute for the Social Sciences]

More policies from UK (20-30%) than Germany & France
- UK: strong centralized funding
- Germany: federal state; strong tradition of independence

UK EPSRC driving this trend

puts requirement on institutions to support data

<RUSSELL GROUP> (high research institutions)

not huge difference b/t in & out of group

WORD#	250	~1000 average	60000

- short strong values weak on details
- longer on requirements

two general policy types

push from newer uni's in this area too

[POLICY CONTENTS]
- ☐ what to keep — various specificity
- ☐ data retention time — when given, 10 years opt for
- ☐ ethical use/reuse
- ☐ access
- ☐ availability (open access)
- ☐ define data (& value)
- ☐ identify university roles/responsibilities
- ☐ DMP requirement — 83% required!
- ☐ who it covers (Staff v. students)
- ☐ ownership — 27% state owned
- ☐ external funding reference
- ☐ subject to review (review after X years)

different work for-hire type conditions

table available via DCC
data available via UKDA

(seeing data under IP policy in a lot of "new" universities)

In Australia, you can copyright a phone-book. Wow.

QUESTIONS

Amy Pienta ICPSR

Looking at ICPSR, do some social and behavioral sciences disciplines share more than others?

SURVEY
- NSF & NIH PIs
- social/behavioral science data
- range of topics on data sharing
  - behavior
  - perceptions
  - barriers
  - publications

879 respondents in this study

See similar results in observed sharing and actual sharing

Data Sharing
- 12% ~~self~~ archived /R
- 45% informal sharing: website, by request
- 43% no sharing

- poly sci/econ more likely to archive than
- physiology/health, informal ~~public~~ preferred
- psychology less likely to share + archive informally
- health sciences less likely to archive
- poly sci/econ more likely to share informally + archive

remove demographics

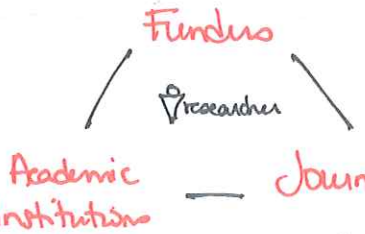
Barriers to Sharing
- ✗ confidentiality
- ✗ lack of time
- ✗ documentation
- ✗ won't understand

different main barriers b/t disciplines

target surveys based on this

Sharing behaviors/views different by discipline

Alexia Katsanidou
GESIS - Leibniz Institute
for the Social Sciences



Big Journal Journal Data Policies

Constrain & push researchers

Political science

higher impact factor
more likely to have policy

BUT researchers don't always comply...

ask authors why shared data?
(same journals as previous study)

questionaire

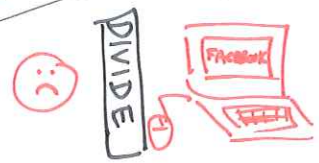
initial results

- ☐ researchers less likely to share if afraid
  - ☐ others may find something new in data
  - ☐ or might be scooped
  - ☐ have to use standard (metadata
- ☐ more likely to share if
  - ☐ data will enhance research
  - ☐ give back to community
  - ☐ suitable repositories available
  - ☐ culture of data sharing

IASSIST Day 2

Curtiss Cobb - Facebook

Measuring the Digital Divide: Using Existing Data Sources and New Data Collection to Understand Between Country Differences



The internet is all about services to people and communities, allowing them to achieve their potential ⇒ extend internet access to developing world

Key facebook principles internet access

The Plan: lower cost x less data
mobile-based internet

of the world's 7 billion people only ~2.7* have access

*mckinsey

Barriers of STOP
- access
- education (literacy)
- gender divide

How do we measure progress?

need to have #'s for background but hard to do in many countries

run into a lot of social issues here

Numbers unknown

Forecasts are black boxes

Unconnected demographics
* poor
* rural
* illiterate
* age

differences

How do we use 3rd party data?

price/availability
source
timeliness
relevance
gut check

DATA EVALUATION

CONTEXT AND CULTURE MATTER

Millions of Facebook users have no idea that they're using the internet

D1 : Data Professionals

Adetoun Oyelude – Data Professionals training challenges in dynamic work environments

University of Ibadan

How do we train ourselves to get ahead?

Interviewed 10 data professionals (different positions)

<self training prominent>

Interviewed results from 10 professionals in Nigeria

CHALLENGES

- funding for training
- infrastructure
- bosses not understanding need for training
- understanding technicalities
- knowing of training opportunities
- lobbying for (more) time for training
- synchronizing time for training online

SOLUTIONS

+ sponsorship
+ social media
+ showcase results
+ reading grey lit
+ learn lobbying
+ grant skills
+ work around obstacles
+ politicize issues
+ sacrifices

WHAT MORE?

- support from above
- professional orgs support
- trainers should train others
- share via social media
- institutional mentoring
- invest in ourselves

use

Construction

arrogance

Michelle Edwards [CISER]

the data librarian lifecycle

~~Stolen~~ Borrowed from Research Data Lifecycle

THE CYCLE:

concept (new sources)

collection (community info)

processing (needs assessment)

archiving (write it down)

distribution (hours + location)

discovery (marketing) analysis

Respond to change ⇒ repurpose

- meet w/ local librarians
- learn what's been done
- create new information/processes
- raise awareness & establish hours
- integrate into community, campus & library
- is it working?

M. Edwards thinking about this w/ respect to recent job move

Guelph → Cornell

Line Pouchard • PURDUE

Comparing policies for open data from publicly ~~available~~ accessible international sources

Had to negotiate agreements b/t provider & Purdue

Concerned about data reuse and access

CAM² project >60,000 webcam streams

Keep data on Amazon servers b/c it's BIG

(don't store stream, just data analysis)

EXAMINE OTHER COUNTRY'S POLICIES

(data protection directive)

often PRIVACY contradicts REUSE

UK has most cameras (estimates that you'll be photographed ~300x/day) in London

US has patchwork of regulations

(different for personal & industry)

click-thru consent

regulations written b/fore arrival of BIG DATA → policies not suited to BIG DATA

BIG DATA ✗ Policy

VARIETY OF POLICY PROVISIONS

E5: Restricted-Use Data Support in Academic Libraries

Jen Darragh - Johns Hopkins
Ryan Womack - Rutgers
Jen Doty - Emory
Jamene Brooks-Kieffer - U. Kansas
Sarah Irwin - Penn State

Physical restriction

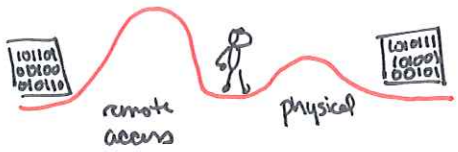
vs

grad students don't have own space

remote access

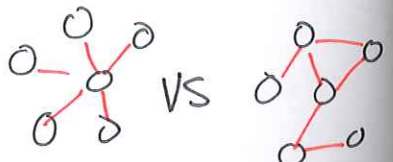
much more technological many issues here

library IT not here



Developing services

- Ryan - student need + extra year-end $ } serendipity
- J. Darragh - limited resources in one area pushed services in other areas
  
  special room for limited access
- Sarah - service limitations due to funding source
- J. Doty - lots of anecdotal evidence (eg. ICPSR application)



overall institutional structure MATTERS

How do researchers store sensitive data?

Survey of institution researchers producing PII (not clinical/medical)

Jamene Brooks-Kieffer

{ contains identifiers or triggers uni classification policy } SENSITIVE DATA

Humanities → Social Science
different practices → Science

Institutional Networked storage not researchers top choice

Researchers use
1. Local computer
2. hard disk
3. DVD/CD/USB
4. Cloud storage
5. institution storage networked
6. institution server
7. institution cloud
8. HPC cluster

researchers using personal accounts here

Researchers want EASY

PROBLEMS WITH NETWORKED STORAGE

- ☐ low capacity, high cost difficult to use
- ☐ insufficient support for sensitive data
- ☐ sync & collaborate difficult
- ☐ incompatible work conditions

Mismatch between needs and services

IF researchers need help w/ own data
THEN need help with sensitive data

J Darragh: library will be your security (physical & electronic)

Can only do so much when there's another person involved

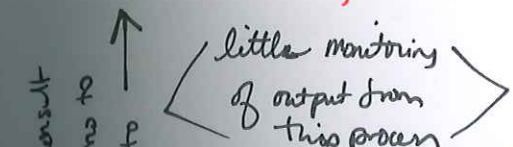
Library can't have total control...

re: security

Ryan Womack:
library workflows & policies for use of restricted space

Sarah Irwin:

Takes time to develop services } working group looking @ cross-campus workflows



What is library's role in this area?

{ training on mitigating disclosure risk } J Darragh

J Doty: sensitive data is part of RDM service, partner with IRB

← Bo Wandschneider

@bjuul:

- user is right in wanting it easy
- institutions need to do better
- Cloud is doing it better than we probably can
- partner w/ IT

stakeholders

- IRB
- IT
- Contracts people
- individual departments
- Campus working groups
- research office

LIABILITY

- issues with admin
- misunderstanding risks / services

+ keep services limited

+ schools know of these issues

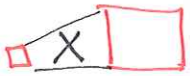
LIBRARY

Solution that helps while university solves (hopefully) larger issues

$ of Services (currently + going forward)

- ☐ J Darragh $10-15k + IT help (room + tech set up) *
- ☐ Ryan W <$2k initial ~$3k for new lock system +

*not scalable, need more manpower



may end up out of library

(+ again, not scalable)

BIG MESSAGE ABOUT SCALE

Is this part of regular data services? Should it be?

J. Darragh

Isn't at Johns Hopkins maybe should be

EDUCATION!

Need to be in addition to storage

Library v. IT role here

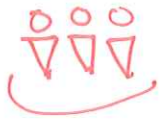
if can't provide service

be CONCEIRGE

stay just a bit ahead of users...

F5: Using DMPs as a Research Tool for Improving data services in academic libraries

Amanda Whitmire – Organ State
Lizzy Rolando – Georgia Tech
Brian Westra – Oregon



bit.ly / dmpresearch

curl for the project

SO MUCH DATA!

researchers don't know how to handle data

libraries can help!

How develop services / understand research needs?

- surveys
- interviews
- review DMPs ← this project

DMP reflects
+ Knowledge
+ Capabilities
+ practices
+ needs

← but need to audit this as verifiable

inform library services

NSF requirements different b/t divisions

general rubric

Need a tool

Using a rubric

GOAL to make rubric accessible to anyone

tested on real DMPs* and entered assessment via Qualtrics

Careful about inter-rater reliability* and other metrics

*inter-class correlation

⇒ used to test rubric to make sure it's clear

* plans from funded projects

Using dmps to look at institutional info and inform data services

LizzyR.

Previous work
- 40% mention IR
- researchers share text

Used plagiarism software

↑ correct misinformation
put more info online (instead of come to library)

Learned more than interviews or surveys

Results of analysis
- 8 plans no data
- 5 plans from math – not producing data?

42 analyzed

∴ good at describing data ~~description~~
∴ bad at describing metadata
∴ mixed re: formats
∴ bad at where data available
∴ mixed at public availability
∴ bad at security measures
? N/A at PII / IP data
∴ bad at reuse
∴ bad at redistribution
∴ bad at production of derivatives
∴ mix at archiving / preservation
∴ unfortunate data sharing places journals?

REDUX
researchers get sharing message
less good at archiving
BUT repository does both!
bad at documentation (not even metadata) sad
share data via journals difficult / more work here

↳ Intervene @ school of math
Improve web presence and add boilerplate

FUTURE STEPS
communicate to campus
repo technical req's
metadata and documentation

<lots of PIs / work here>

Brian Westra

NSF biology guidance
data AND metadata
appear together

assess ⇒ Post-award

now moving to annual report

BIO sharing
- personal website
- on request
- named data centers
- journals

"What is data?"
How do you publish it?

still an issue

CHEM Guidance language
- peer-reviewed journals
- personal website
- data repository

*this is me shaking my fist at my fellow chemists

Named data centers

crystallographic data big one

NEED MORE DATACENTERS IN CHEMISTRY

⇑ actually are sharing "data" as journal supplements

Lots more focus on by-request <collective sigh>
actually in guidance language

Plenary 3

The Politics of Open Data

Andrew Johnson
- Minneapolis City Council -

Johnson:
- Data analytics
- web development
- entrepeneur
- interest in politics

Bill listing where tax money goes ⇐ "What we Pay For" website

TRANSPARENCY

People like this

ran for city council

HE WON!

DATA VIZ CHALLENGE via google

IT subcommittee
Budget directive
work w/ CIO

"it will take 4 years to do"

FEAR of OPEN DATA
- open data will make us look bad
- without context, data could be misinterpretted
- people could do anything with this data!
- will cost us millions!
- geeks will have an unfair advantage

↑ digital divide in city

BENEFITS
+ public trust
+ shifts relationship
+ reduces requests
+ better data quality
+ empowers staff
+ empowers community to further city goals

seen examples of this already

POLICY
PORTAL
CULTURE

harder

challenges
- legacy systems
- open data v. open process
- prioritizing data
- portal politics
- political process

GOALS
empower leaders
gain buy-in

success stories
more pressure to be open

dataset can make

last-minute attempt to make policy optional

POLITICS ↗

outreach to stakeholders

↑ restaurants: food safety info going online

NEXT STEPS
- ☐ compliance report
- ☐ revisit portal
- ☐ events for coders
- ☐ continue culture change
- ☐ revisit policy

add open policies into purchasing

16th city w/ open data policy

G2: Planning RDM Services

Kate McNeill
MIT Libraries

RDM Meets Open Access
(RDM = Open Access)

RDM ↔ OA

- fosters open access to information
- focus on information producers
- Compliance w/ funder reqts
- publishes policies

DIFFERENCES

RDM: complexities; overlap w/ other areas; whole lifecycle; complex collaborations

OA: copyright complications; end of lifecycle; more familiar

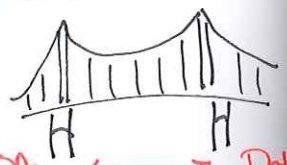
In MIT Libraries...

"data is specialized service"

Using the team approach

OA } Office of Scholarly publishing, copyright, & licensing

OA Logo! where's our RDM Logo?



make services work together

use bottom-up approach but make structural

Working jointly

- Communication
  - in library
  - on campus
- increased awareness
- synergies

JOINT GROUP
Coordinate but not provide services

High-priority Library projects → + OSTP Memo + Share personnel

meets bi-weekly

UNCERTAIN

who does what
authority
representing other services
strategic planning?
skilled @ collaboration?

COLLABORATE

Partnerships in a Data Managment Village

Alicia Hofelich Mohr
Lisa Johnston
Thomas Lindsay
U Minn

Researchers get alot of services along data journey

need to add data steps to research process

how do we reach researchers more?

IT TAKES A VILLAGE

where are our village people?

lots of places in uni to connect

Connectivity & Awareness

start to connect engaged partners

more beyond network to partners

U MINN CAMPUS

- STRATEGIC FOCUS
  - policy
- COLLABORATION OPPORTUNITIES
  - community of practice
  - RDM Library group
  - design thinking workshops

Tom L.

[College of Liberal arts + Libraries]

- DMP Training
- Student workshops
- work in methods course
  - mix breadth / depth
- IRB Ethics ~~training~~ consults
- Coordinate b/t partners
- documentation templates
- de-identification
- DRum repository
- Curation model (distributed)
- enable / track reuse

how long is the new forever?

re: data retention

Lisa J.

how to prep data for sharing

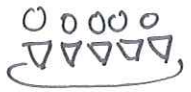
DATA VILLAGE

- doesn't happen overnight, work in progress!

think beyond each village

Carol Perry
University of Guelph

data management on a shoestring budget



} training from various sources

5 members of data + schol comm team

REPOSITORY PLATFORM
Dataverse + DataCite

SCALE
- focus local
- not huge branding
- pilot
- small datasets

hard $ manage staffing

COLLABORATE
+ library
+ office of research
+ computing & communication services

DEVELOPED SUPPORT MATERIAL
website
modules
workshops

RDM Services
- Environmental scan of services
  - models
  - lit review
  - brief faculty survey
  - interview deans / dept Chairs
- DMP support
- Support for new funder requirements
- Data consultations
- Data curation

OUTREACH
Outside funding for repository
research data canada seminar
advise another grant agency on setting up reps
consortial ~~memberships~~ models
academic of DMP template
~~shortages~~

Planning for upcoming DMP requirements