



RDA/WDS Publishing Data IGs and WGs

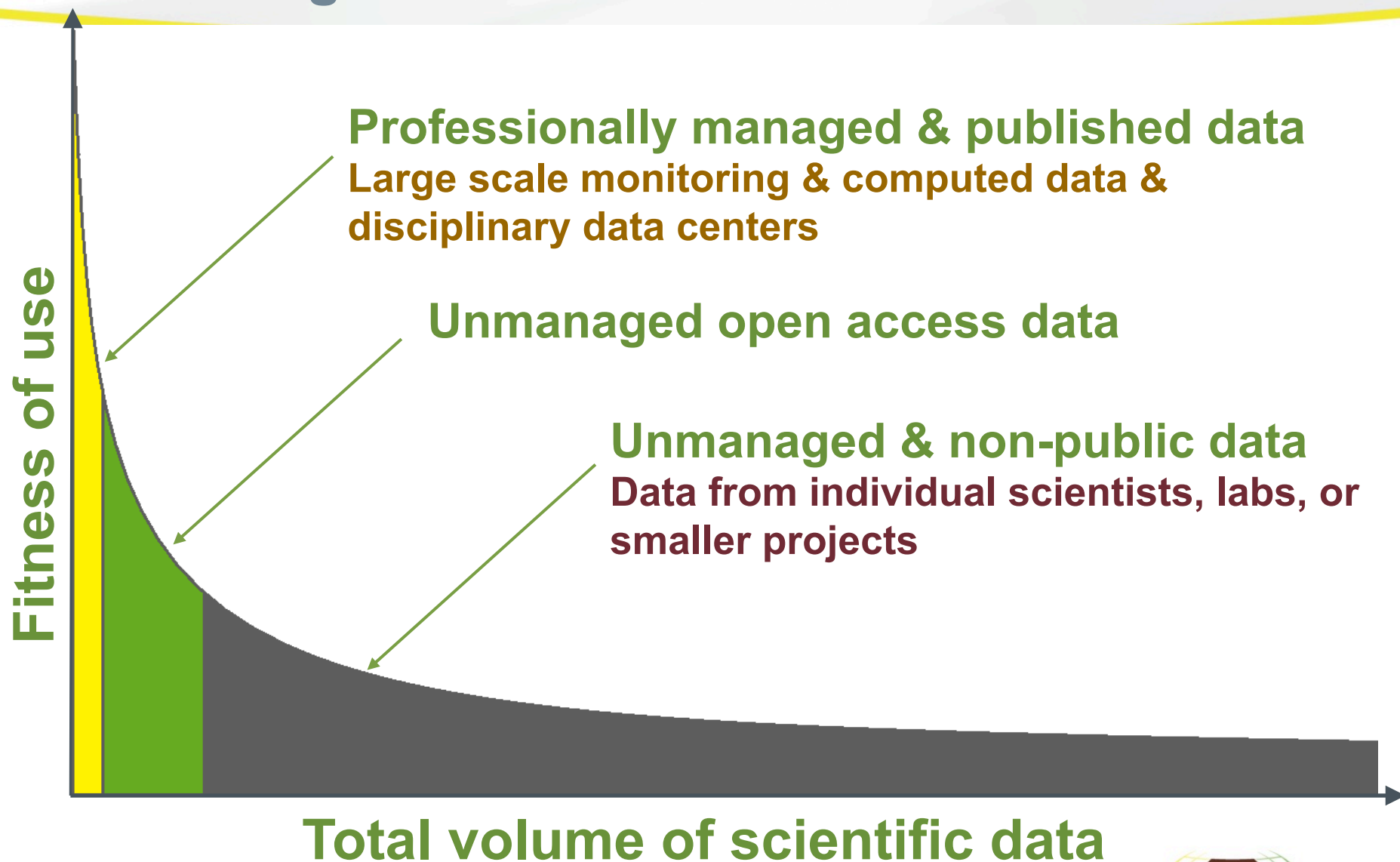
**3rd Working Group Collaboration Meeting,
June 2015, Karlsruhe**

research data sharing without barriers
rd-alliance.org

RDA-WDS Data Publishing IG

| | |
|---|-----------------|
| Theodora Bloom (BMJ) | - Workflows |
| Adrian Burton (ANDS) | - Services |
| <u>Sarah Callaghan (BADC)</u> | - Bibliometrics |
| Todd Carpenter (NISO) | - Bibliometrics |
| Sünje Dallmeier-Thiessen (CERN) | - Workflows |
| Michael Diepenbroek (PANGAEA) | |
| Ingrid Dillo (DANS) | – Cost Recovery |
| Simon Hodson (CODATA) | – Cost Recovery |
| Hylke Koers (Elsevier) | - Services |
| John Kratz (CDL) | - Bibliometrics |
| Kerstin Lehnert (IEDA) | - Bibliometrics |
| Mustapha Mokrane (ICSU-WDS) | |
| Eefke Smit (STM) | |
| Jonathan Tedds (D2K/University of Leicester) | |

The Long Tail of Data

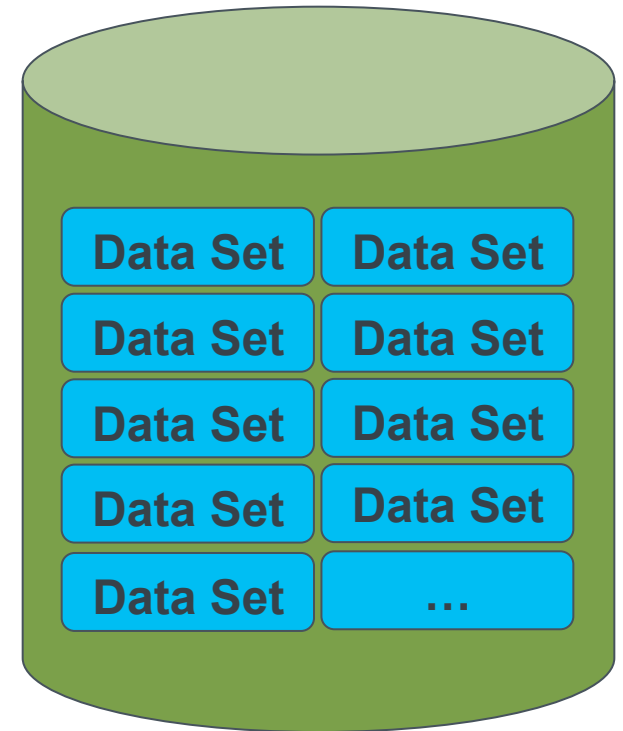


Data publication - prerequisites

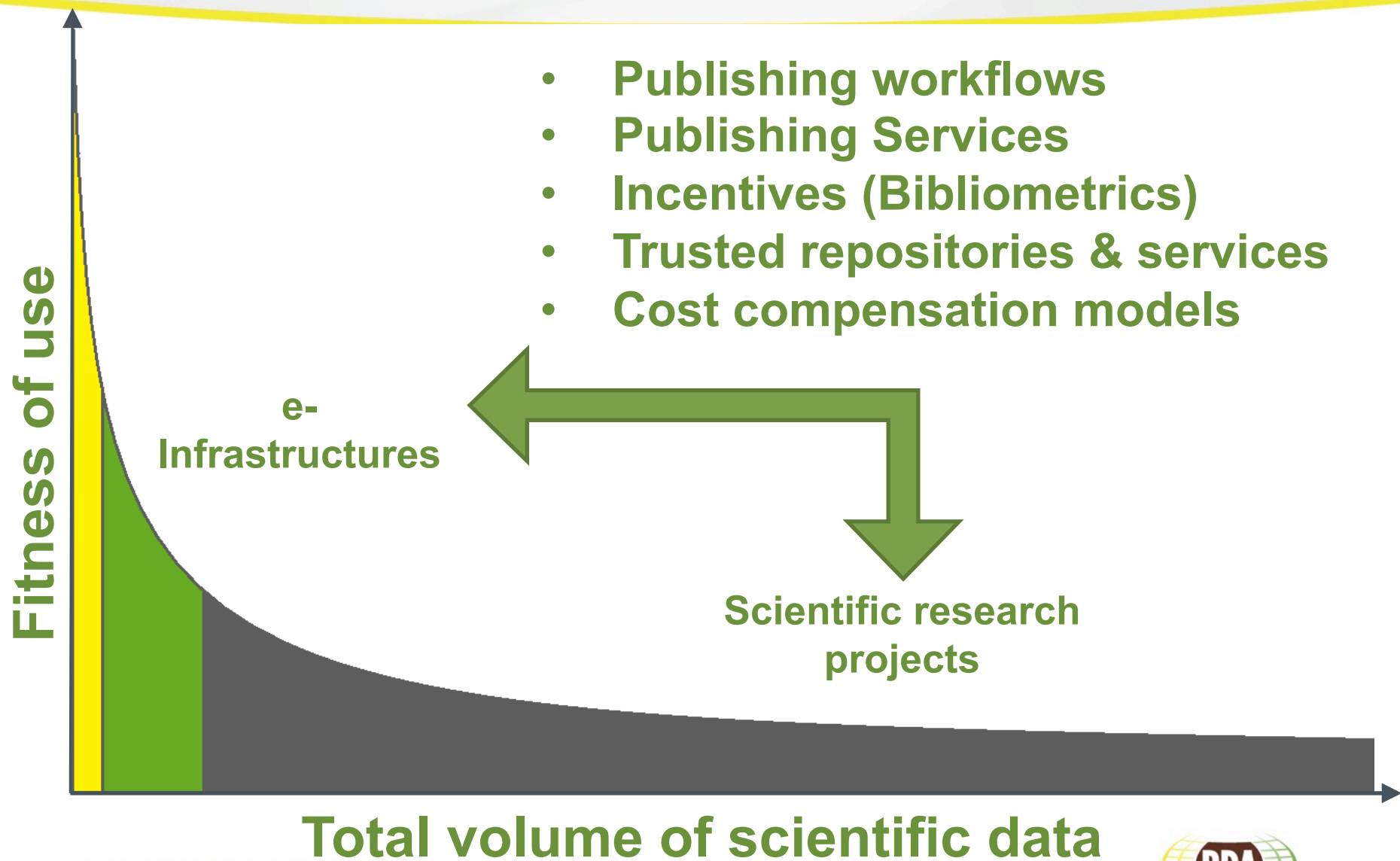
4

OECD principles and guidelines for access to research data (2007)

- Licenses & persistent identification
- Quality
 - ✓ QA/QC -> review procedures
- Efficiency
 - ✓ (Meta)data & interoperability standards (machine readable)
- **FITNESS OF USE!**



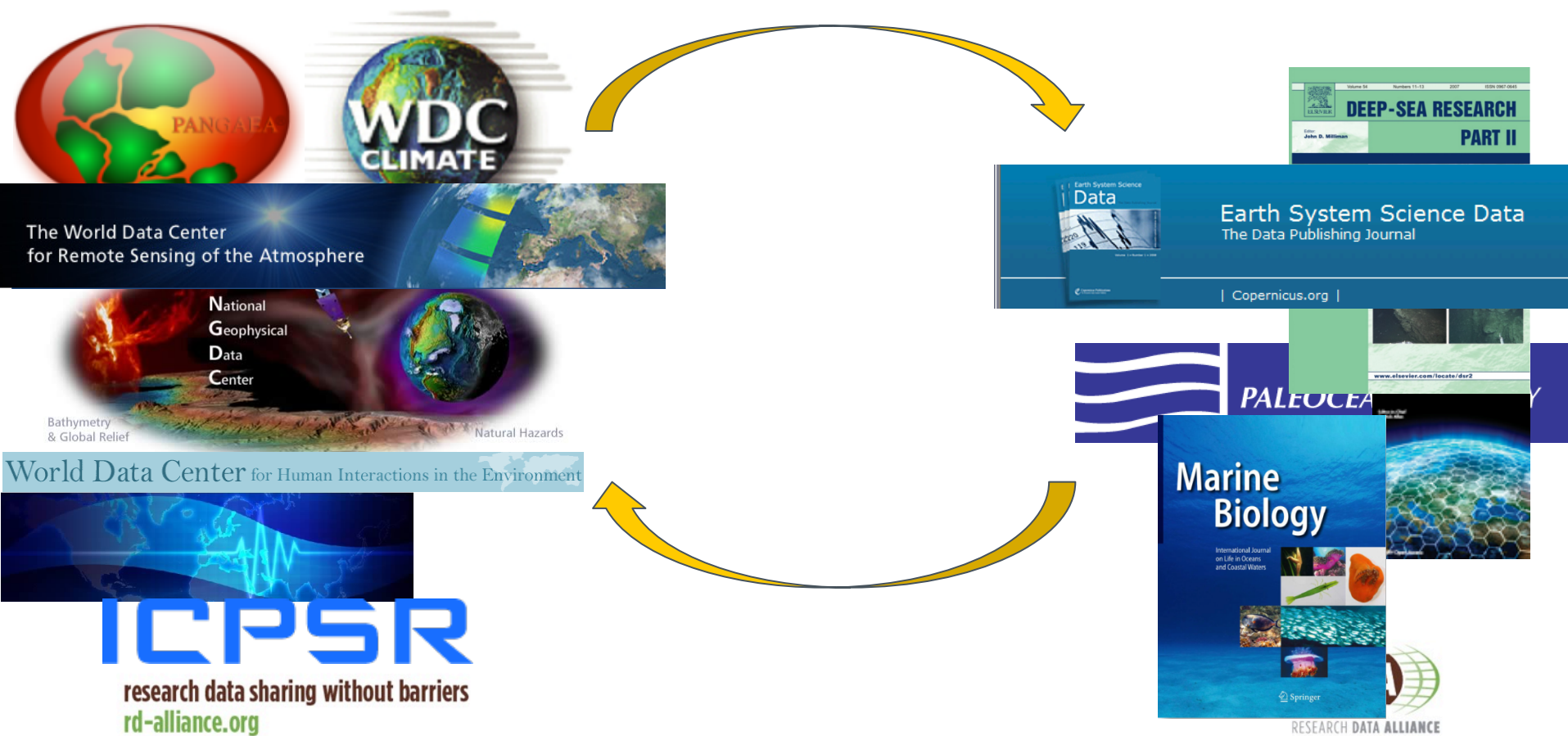
Bridging domains



Collaboration between data archives & science journals

6

- ✓ linking editorial workflows
- ✓ linking services



Consortium

- Research facilities
- Data repositories
- Universities
- Libraries
- Industry



RDA-WDS Data Publishing IG

Coordinated approach



- Collaborative tools
 - Regular web-meetings (2-6 weeks)
 - Common inventory on initiatives & projects
 - Bibliography of articles related to data publishing
 - All documents on Google Drive (open access) -> RDA website
- Webinars
 - 2014-05: Publishers
 - 2014-10: Data centers & services
- Coordination with related groups
 - data citation, Force11, RMAP, NDS, certification, PIDs
- Funding proposals (H2020)
 - 2014: THOR (ORCID, DataCite, Data Centers)
 - 2015: Call GARRI-4 meets the objectives of the Data publishing WGs
 - Presentation in the H2020 e-Infrastructure meeting



Data Publication Services WG

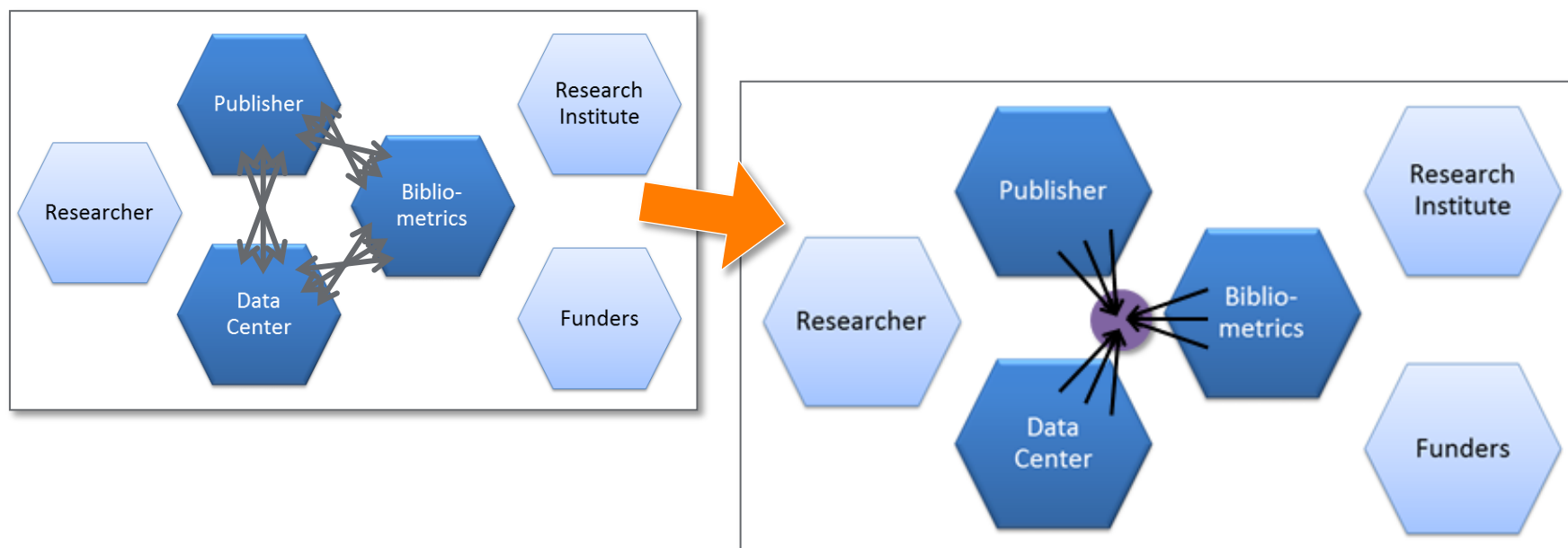
Chairs: Hylke Koers, Adrian Burton

research data sharing without barriers
rd-alliance.org

Introducing the joint ICSU-WDS / RDA Working Group

“Data Publication Services”

The challenge in today's data publishing landscape: how do we move from a plethora of bilateral arrangements to a one-for-all service model ?

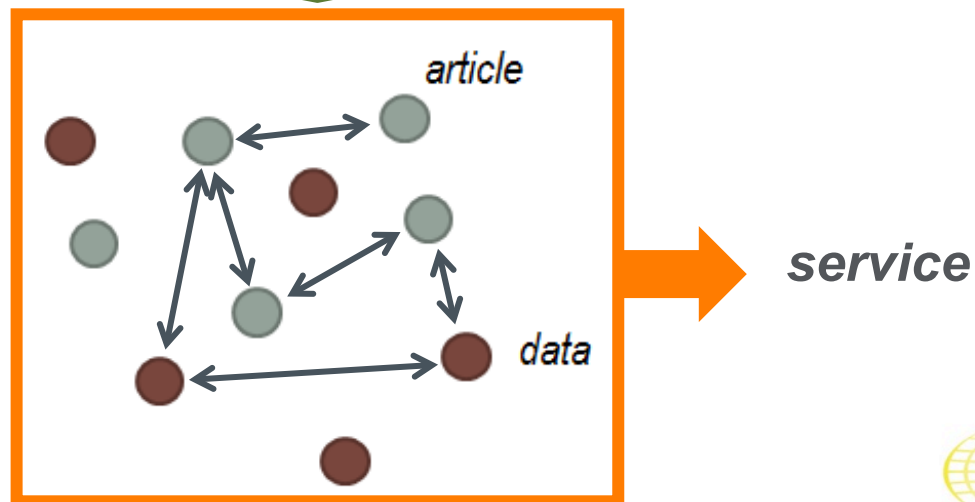
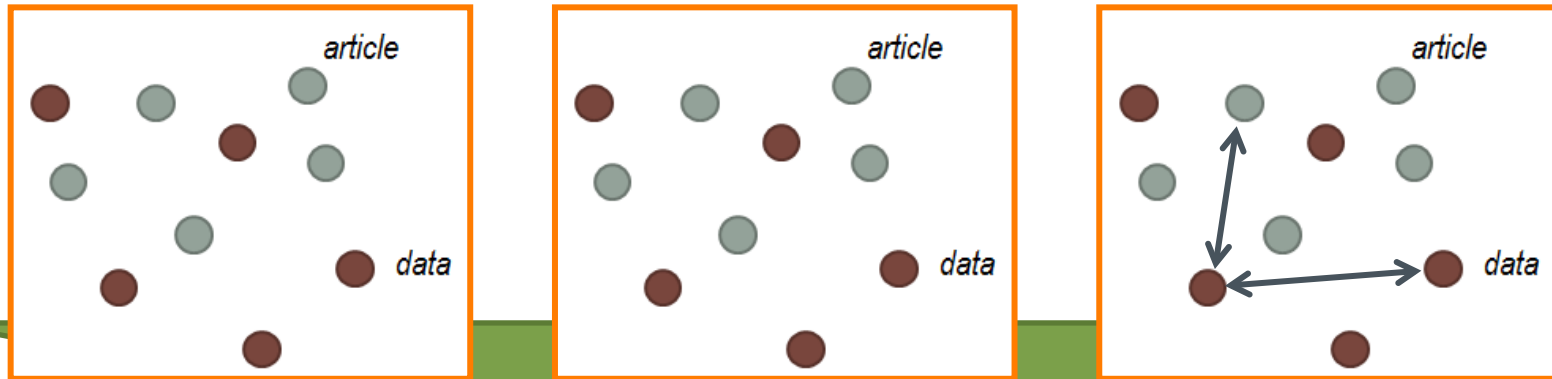


- Increase interoperability
- Decrease systemic inefficiencies
- Power new tools and functionalities to the benefit of researchers

Primary Focus: One-for-all, Article – Data cross-referencing service

- Given article A, what relevant data D exists – and vice versa
- Additional metadata about the nature of the relationship, e.g. supplementary data, related data, etc.
- Additional metadata for article and/or data set

Or.. “it’s all about connecting the dots”



Guiding principles have been ratified

1. The WG will create a **common, open, universal cross-referencing service** for research articles and research data.
2. The WG functions as an open organization in which the key stakeholders are represented. Participation is open to any organization that has an interest in scholarly communication or research data.
3. The WG has a **global** and **multidisciplinary** scope.
4. The WG strives for maximal transparency and all WG meeting notes will be made publicly available.
5. All software developed by the WG will be made openly available under terms that guarantee public access and enable reuse.
6. The corpus of article-data links that will be assembled by the WG will be made openly available under terms that guarantee public access and enable reuse.
7. The cross-referencing service delivered by the WG will be **open and free of charge**. Terms for access to the service will be openly available and non-discriminatory.
8. The WG will advise on the long-term (beyond the lifetime of the WG) management and governance structure, according to the principle that the service shall be governed by a body with representatives from key stakeholder groups including both for-profit and not-for-profit organizations.
9. The WG will advise on long-term (beyond the lifetime of the WG) sustainability models, according that the service will be run on a not-for-profit basis. This may include a hybrid model of both free and paid-for services for cost recovery.

As a researcher looking for data or articles (“consumer”), I want to:

- **Search for relevant articles using a data set identifier.**
- **Search for relevant data sets using an article identifier (DOI).**
- Search for related data sets using a data identifier
- Search for articles, data and links using basic metadata like titles, authors, funder, facility

And when I find links, I want to:

- **Be able to inspect what the nature of an article-data link is**
- Be able to inspect basic information with each link returned (title, author, etc)
- Be able to find out what the origin of the link is

As a researcher publishing articles and/or data (“producer”), I want to:

- Rely on publishers or data centers to link up my articles and data as appropriate
- Be able to deposit article-data links myself

And when links are created, I want to:

- Honor embargo periods for the publication of data and/or article

User stories: research institutes, data centers, publishers, service providers, etc. ¹⁵

As a data center, publisher, or provider of bibliographic information, I want to:

- Construct overviews of all articles linked to one or a collection of data sets based on flexible search queries (*see researcher use cases*)
- Construct overviews of all data sets linked to one or a collection of articles based on flexible search queries (*see researcher use cases*)

And for those overviews, I want to:

- Be able to filter on the nature of the article-data link
- Search using all (or a selection) of my data sets or articles
- **Have programmatic (API) access to the links in such a way that I can incorporate them on my platform.**

As a research institute, funder, provider of bibliographical information, or research portal / aggregator, I want to:

- Get a comprehensive overview of research output including both articles and data

As an article-data link contributor, I want to:

- **Deposit a batch of article-data links in the way that is easy and compatible with my systems**
- Get an overview of what links I have contributed
- Be able to edit or remove links earlier deposited by me
- Get an overview of how often the article-data links that I have contributed show up in search results

Technical requirements to do this all well

- Deduplication of records coming in from different sources
- Deal with data that's versioned
- Connecting to ORCID for authorship information
- Deliver high-performance services (API's) for machine interfacing
- Interfacing with Google et al. for SEO

Essential value elements for a great cross-referencing service

17

As a data center, publisher, or provider of bibliographic information, I want to:

- Construct overviews of all articles linked to one or a collection of data sets based on flexible search queries (see *researcher use cases*)
- Construct overviews of all data sets linked to one or a collection of articles based on flexible search queries (see *researcher use cases*)

And for those overviews, I want to:

- Be able to filter on the nature of the article-data link
- Search using all (or a selection) of my data sets or articles
- Have programmatic (API) access to the links in such a way that I can incorporate them on my platform.

As a research institute, funding agency, or research portal / aggregator, I want to:

- Get a comprehensive overview of all articles and data

As an article-data provider, I want to:

- Deposit a batch of article-data links in the way that is easy and compatible with my systems
- Get an overview of what links I have contributed
- Be able to edit or remove links
- Get an overview of what links I have contributed show up in search results

Technical requirements

- Deduplication of links coming in from different sources
- Deal with data that's versioned
- Connecting to ORCID for authorship information
- Deliver high-performance services (API's) for machine interfacing
- Interfacing with Google et al. for SEO

As a researcher looking for data or articles ("consumer"), I want to:

- Search for relevant articles using a data set identifier.
- Search for relevant data sets using an article identifier (DOI).
- Search for related data sets using a data identifier
- Search for articles, data and links using basic metadata like titles, authors, funder,

And for those links, I want to:

- Be able to inspect what links are available
- Be able to inspect what links are available (e.g., title, author, etc)
- Be able to inspect what links are available

As a researcher looking for data or articles ("producer"), I want to:

- Rely on public data centers to link up my articles and data as appropriate
- Be able to deposit article-data links myself

when links are created, I want to:

- Honor embargo periods for the publication of my article

Flexibility in/out

Provenance

Comprehensiveness

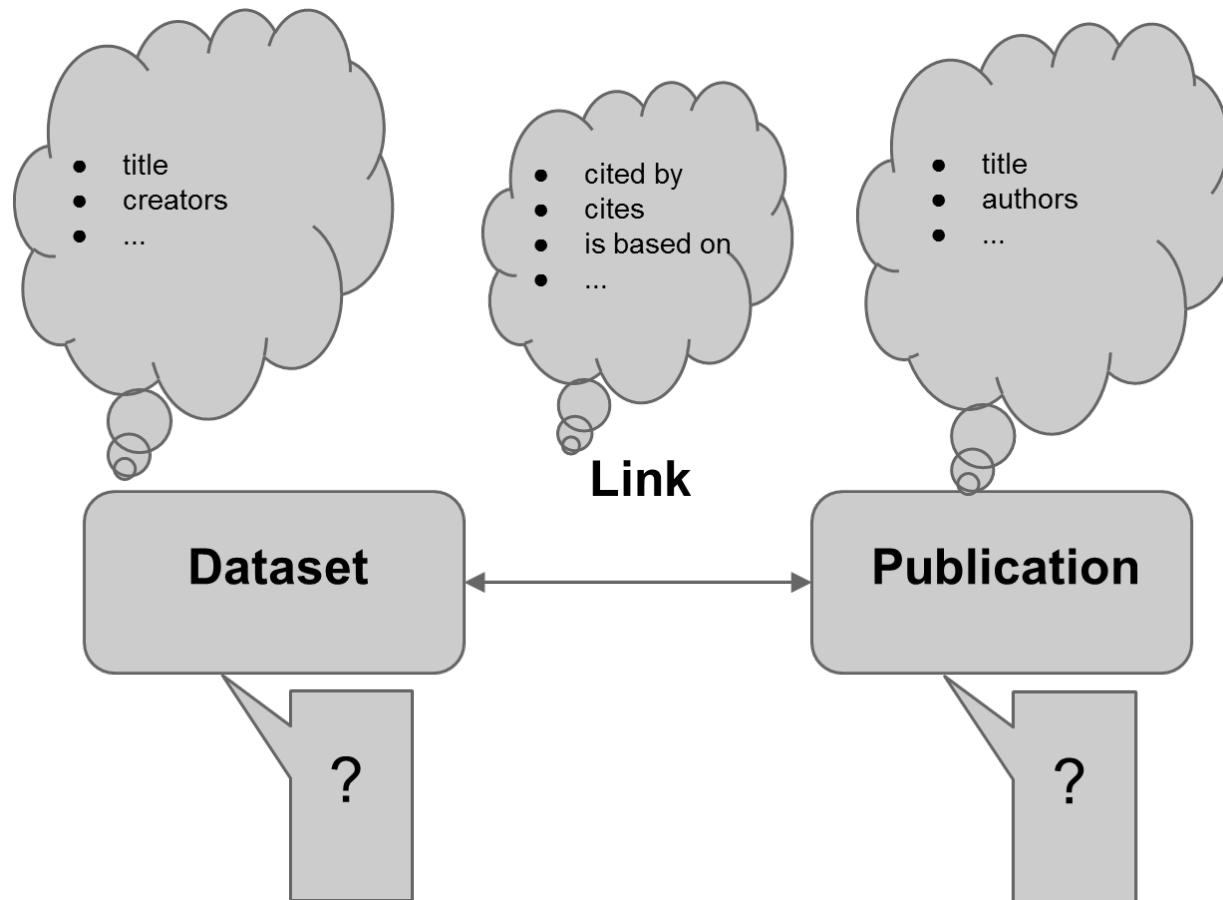
Discoverability

Quality

Metadata

Schema (metadata + information model)

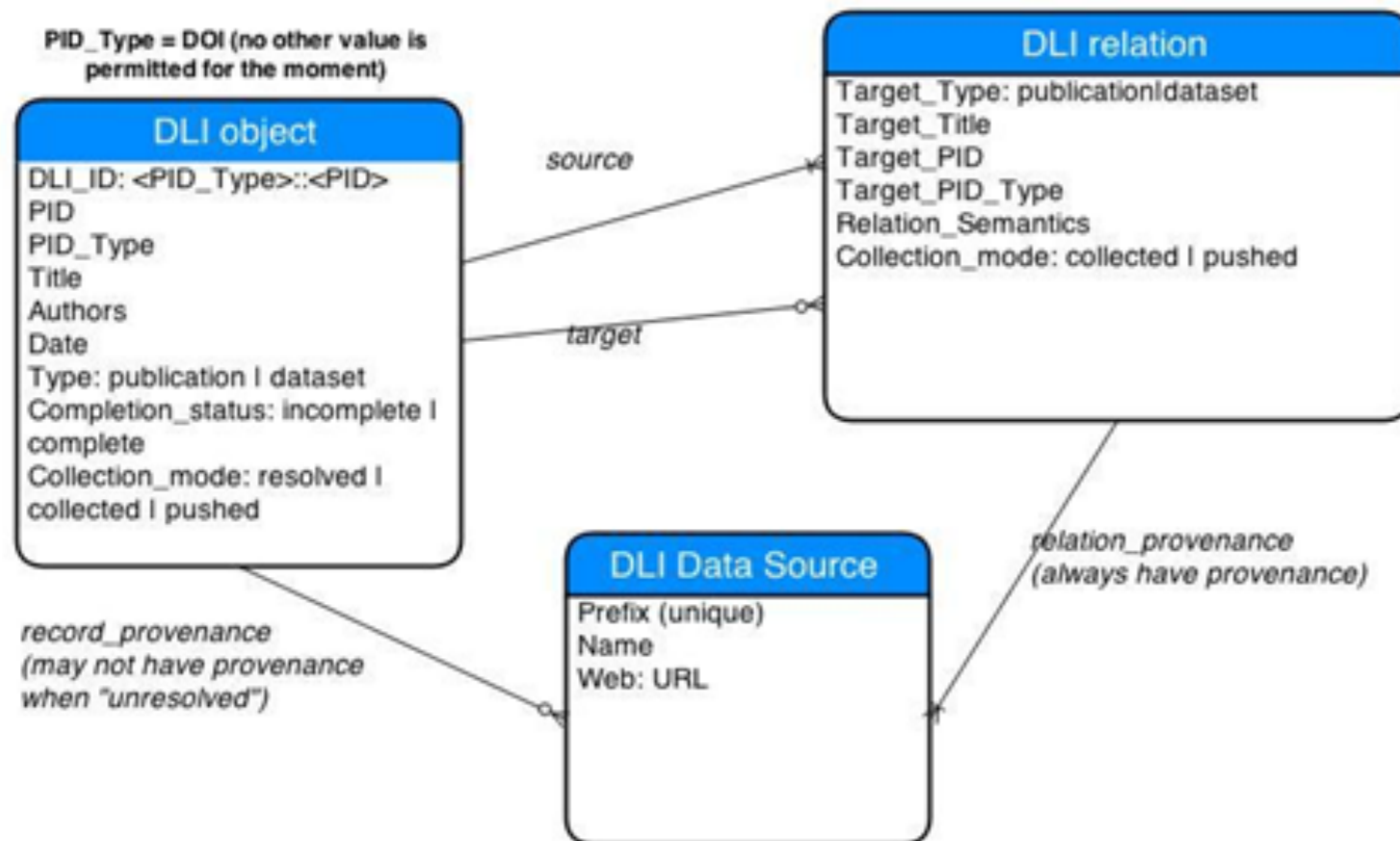
18



(from P4 presentation)

Schema (metadata + information model)

19



Schema (metadata + information model)

20

Discoverability

Metadata

Provenance

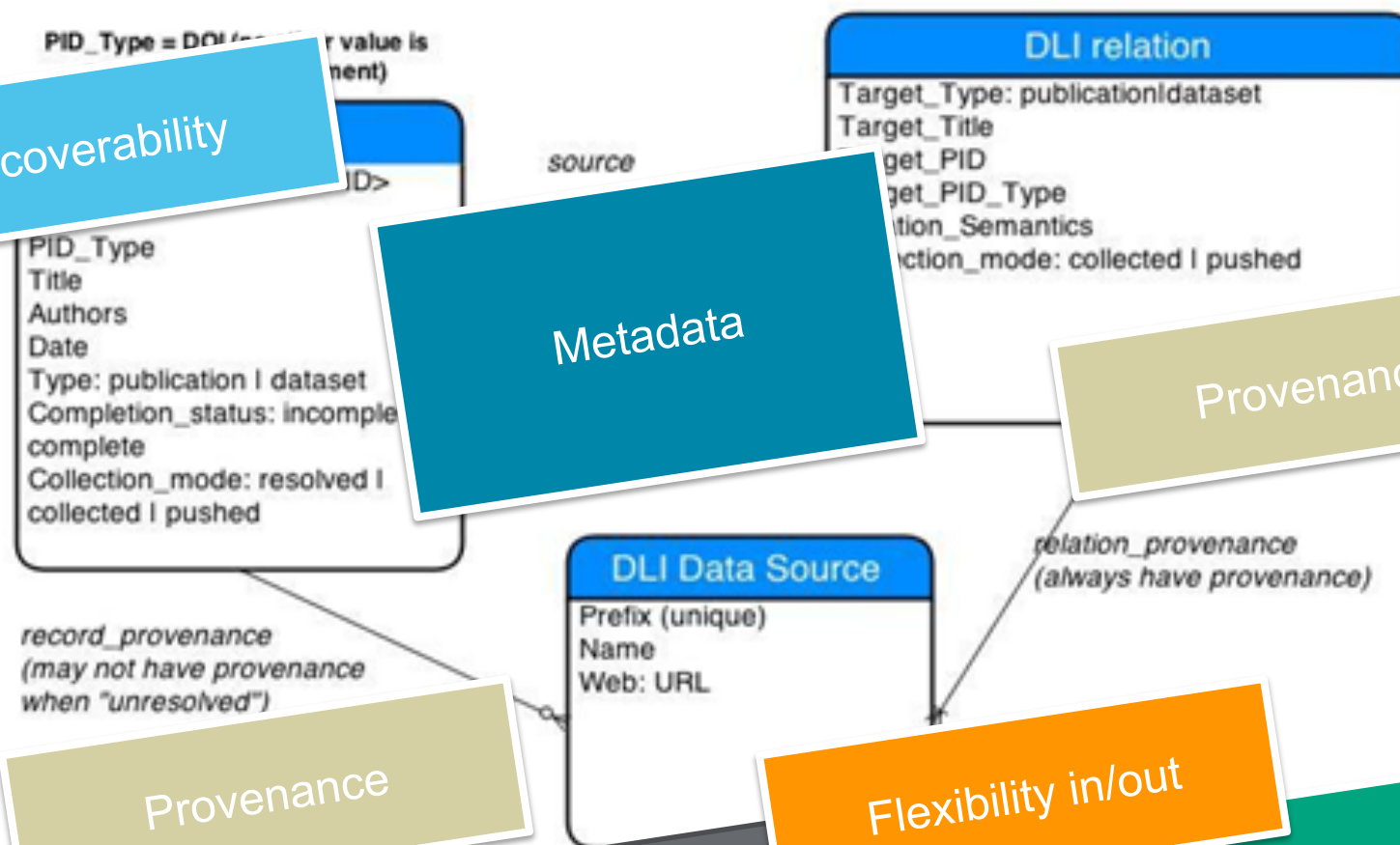
*record_provenance
(may not have provenance
when "unresolved")*

Provenance

Quality

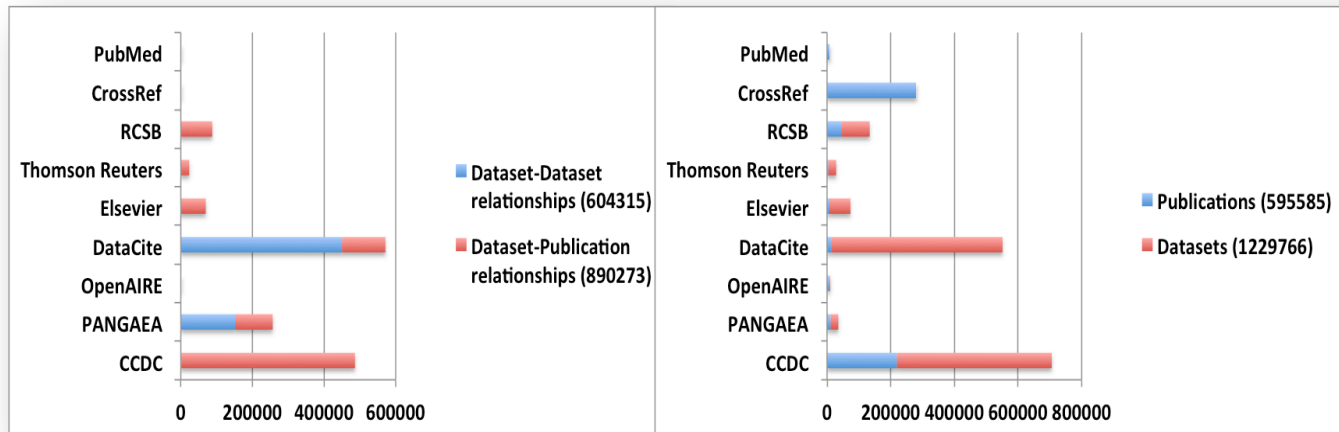
Flexibility in/out

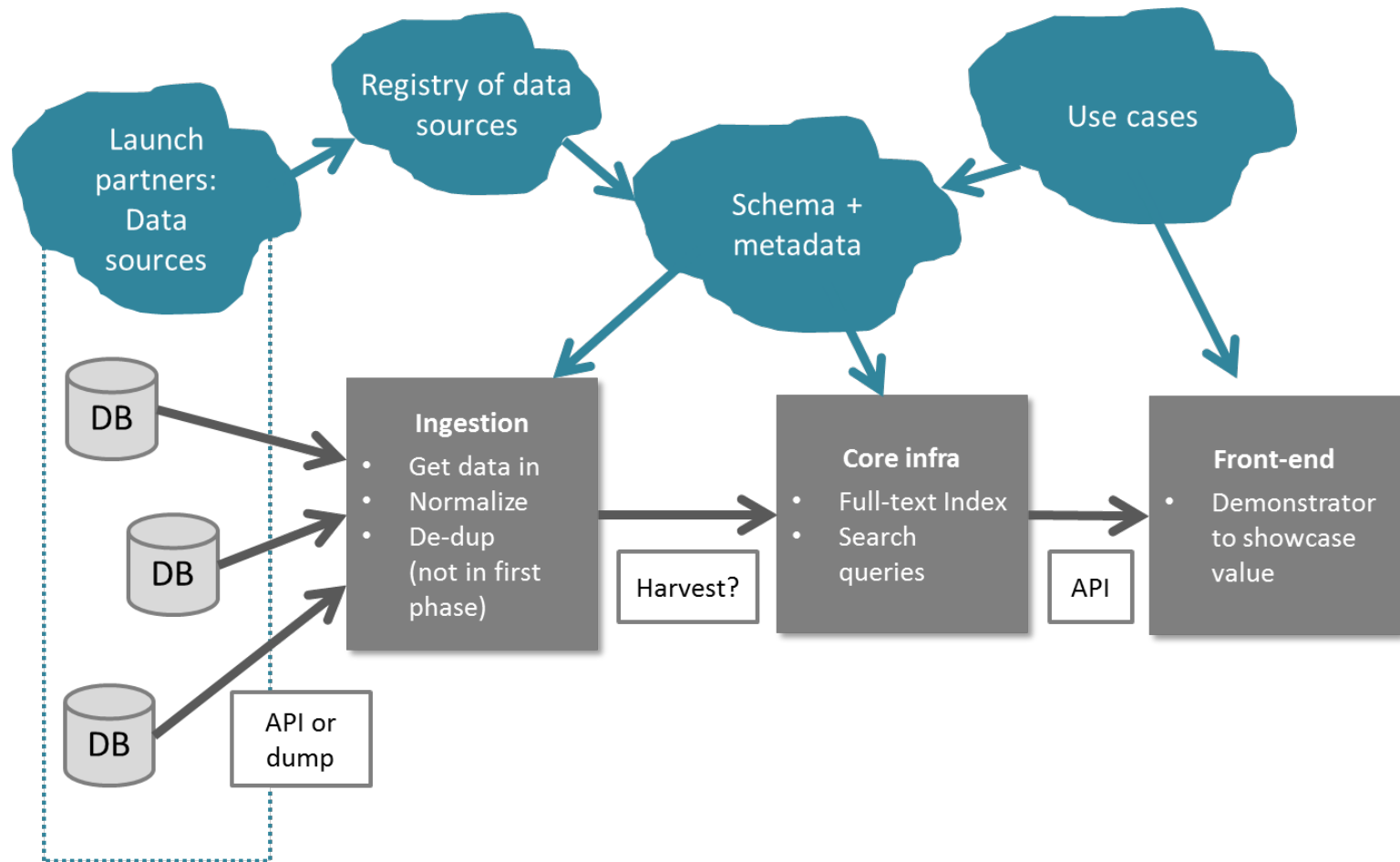
Comprehensiveness



Status

- 890,273 article/data links from 9 sources – more under way
- Metadata schema to capture:
 - Link relationship
 - Provenance
 - Metadata for articles and data where possible
- Operational ingestion, harmonization, completion workflows





Technical demonstrator: 1st prototype results

23

GENERAL INFORMATION

| | | |
|------------------------|--|---------------------------------|
| Title | TPPII, MYBBP1A and CDK2 form a protein-protein interaction network | |
| Object Identifier | 10.1016/j.abb.2014.09.017 | original record |
| Object Identifier Type | DOI | |
| Author | Nahálková | |
| Author | Tomkinson | |
| Collected from | • crossref | |

CONNECTED ENTITIES

IsReferencedBy :

[ncbi-n:AL024407](#)

[More info](#)

PID Type: accession number
Entity Type dataset
Relation Provenance: • elsevier

IsReferencedBy :

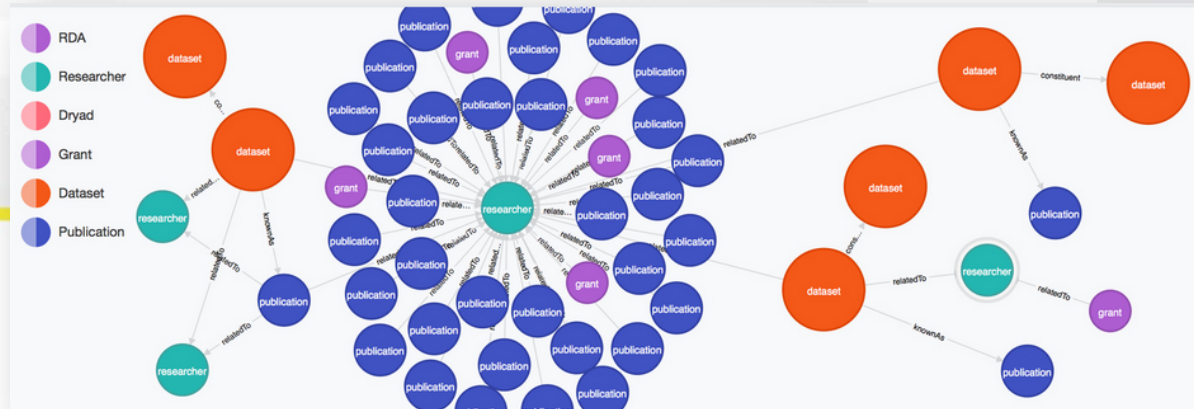
[ncbi-n:AU019902](#)

[More info](#)

[Live demo](#)

- **RMap**
- National Data Service
- BioCaddie
- Open Science Foundation
- CrossRef – DataCite linking service
- **RDA WG Data Description Registry Interoperability**
- *(and more!)*

Next steps



- Sign up more contributors and supporting organizations
- Extend corpus of article/data links
- Connect ingestion system with search-optimized database, API's
- Connect to RD-Switchboard.org for visualization tools
- Complete prototype phase 2 and 3
- Optimise information model (based on prototype)
- Attract more article-data link contributors (and other contributions too!)
- Formalise relationship with DataCite-CrossRef service
- Outreach, webinars, presentations
- Collaboration



RESEARCH DATA ALLIANCE



WORLD DATA SYSTEM

RDA/WDS IG Cost Recovery Models

Chairs: Ingrid Dillo, Simon Hodson

research data sharing without barriers
rd-alliance.org

- Basic 'structural' funding of data infrastructure may not keep pace with increasing costs
- Need to consider alternative cost recovery options, innovative solutions and a diversification of revenue streams
- Not just who will pay for public access to research data, but how these payments will be made



“Despite the growing demand for data sharing and access, domain repositories face an uncertain financial future in the United States”

Objectives and Deliverables



- **A contribution to strategic thinking** on cost recovery by conducting research to understand current and possible cost recovery strategies for data repositories
- What **business models** exist for the data centres that provide public access, what innovative ideas are being developed and what options are open to diversify their income?
- **Report providing conclusions and recommendations** about the appropriateness of different cost recovery models to different situations and the potential of data publication initiatives fitting into a cost recovery strategy.

Where are we now?

29

RDA P5 San Diego

Preliminary Survey results

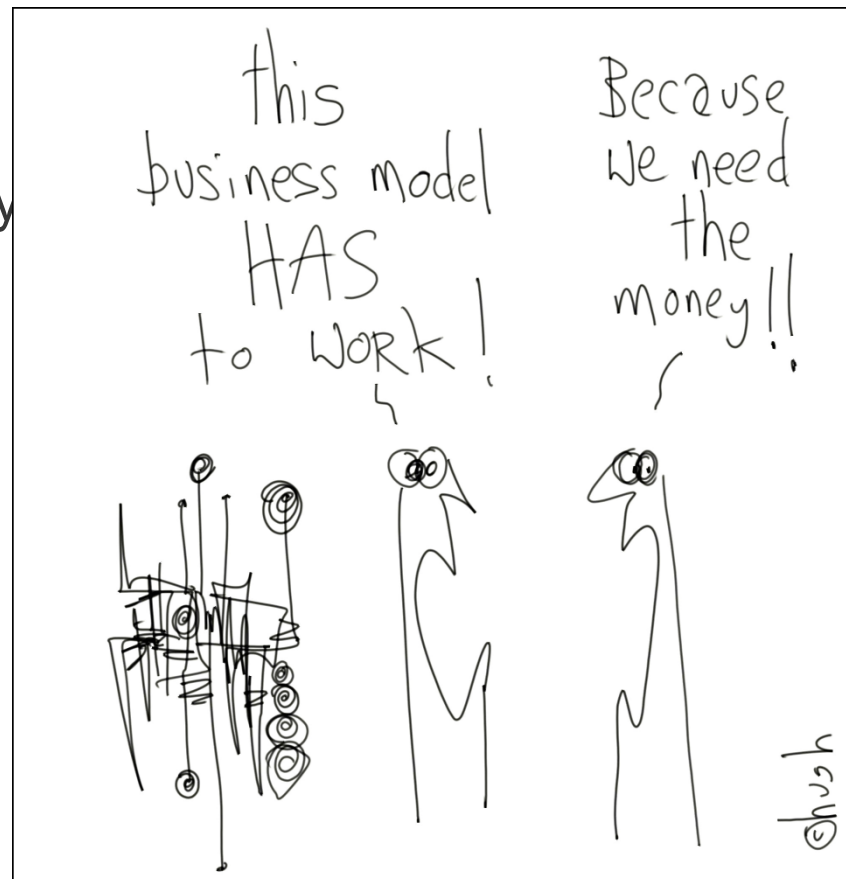
- 23 Repositories interviewed
- Done by volunteers over phone/Sky
- In-depth interviews with script

RDA P6 Paris

- Draft survey report
- Additional case studies
- Discussions with funders and data centres

RDA P7 US?

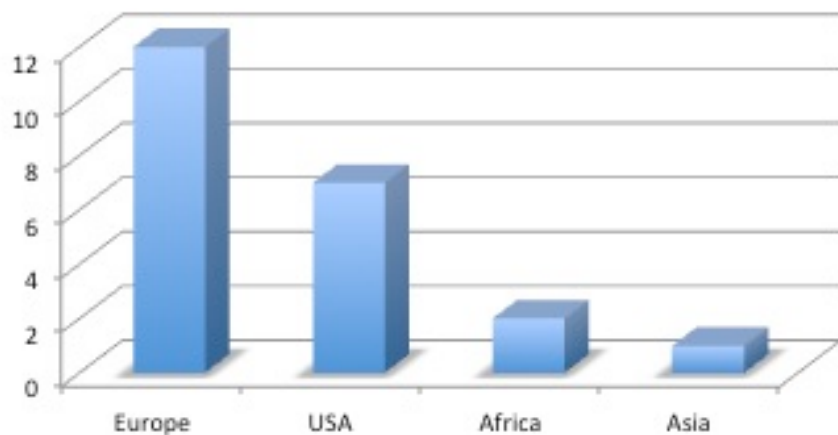
Final conclusions and recommendations



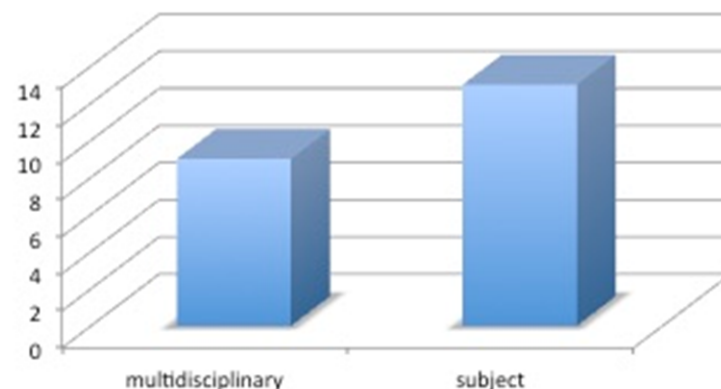
Preliminary Survey Results

30

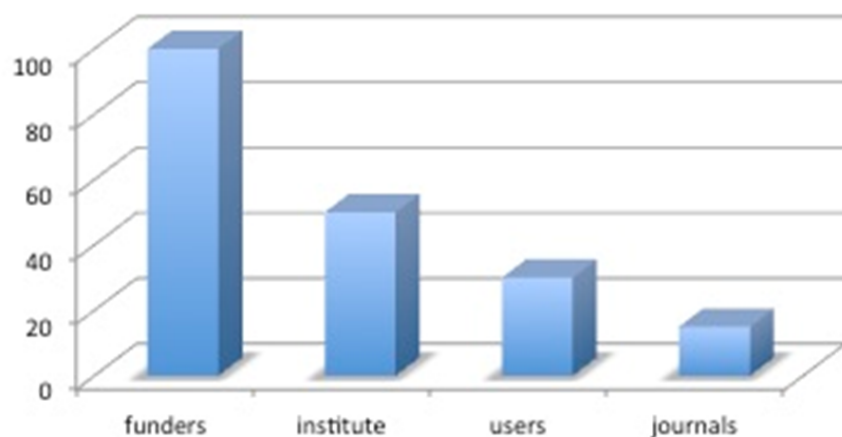
Geographical spread of repositories



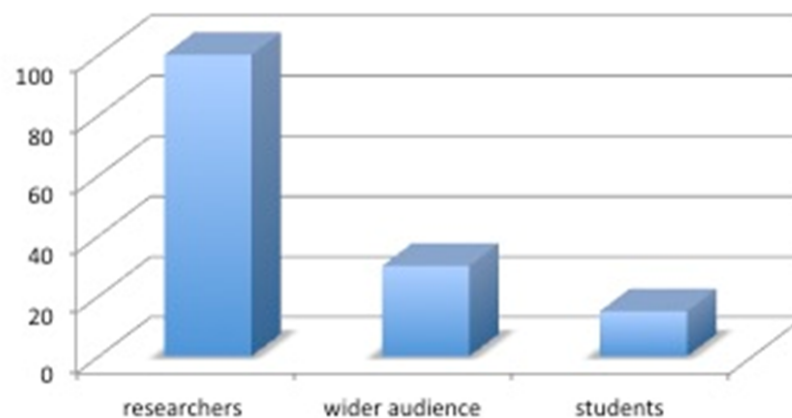
Type of Repository



Repository stakeholders (in %)



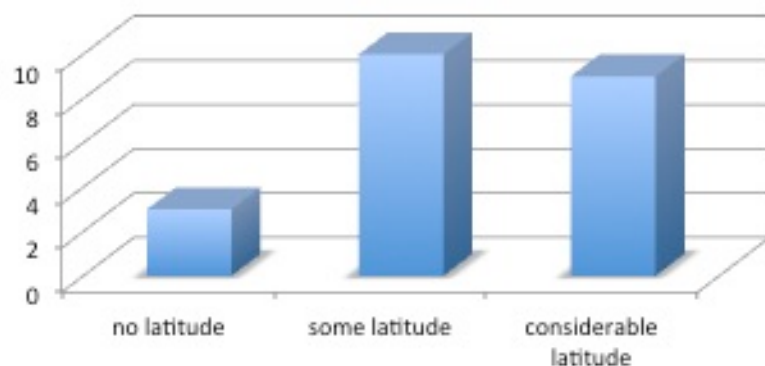
Repository Target Audience (in %)



Preliminary Survey Results

31

Latitude to determine mission and collection policy



Curation levels



Preliminary Survey Results: funding sources

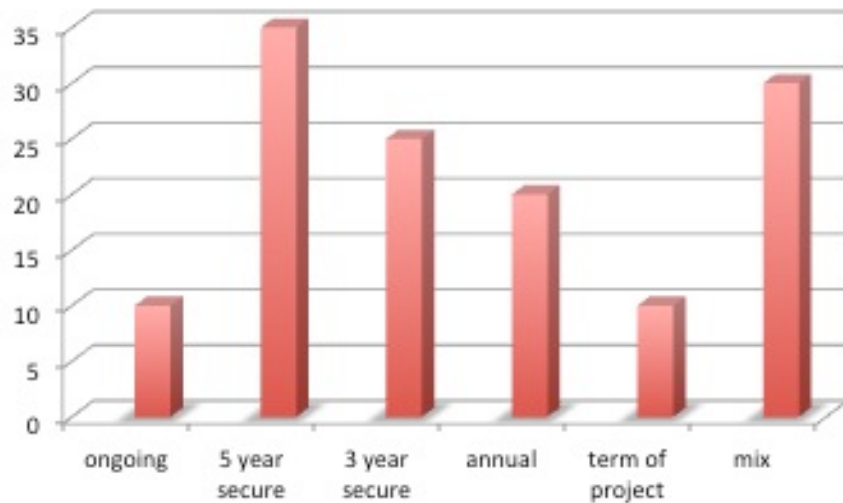
32

Multiple sources of funding:

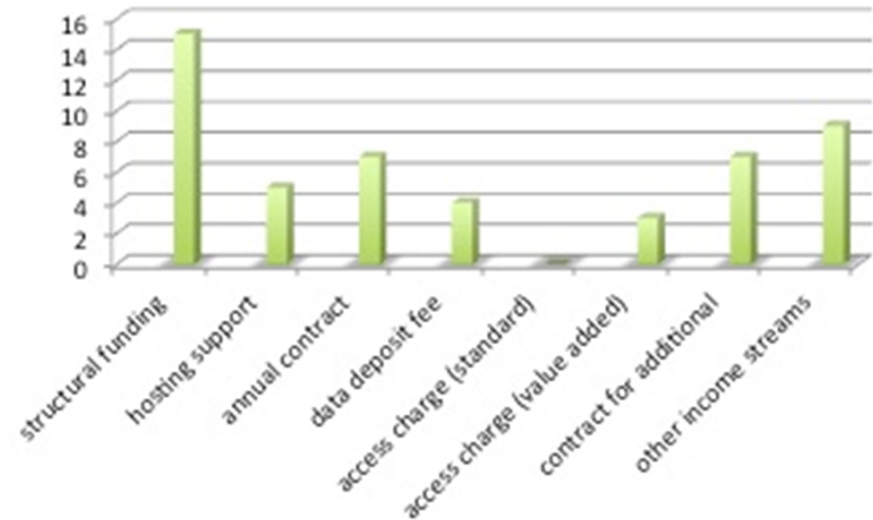
- the (group of) institutes they belong to: 60 – 65% of the repositories
- Government funds and other research funders: 50% of the repositories
- Specific project grants: 50% of the repositories
- Deposit fees: 25% of the repositories
- Annual member fees: 25%

Is the current stream sufficient for the future?

- 60% yes, 30% no, 10% maybe.



Term of funding for the main income stream (in %)



Income streams in absolute numbers of repositories

Structural Funding

- Sometimes covers more services than pure curation (e.g. value-added services, analysis tools, training etc).
- Structural funding works very well for some data centres: many reported that for now funding was adequate for services provided, but not all..
- Others reported that there is a gap in funding which is completed by grant funding.
- General concern that the structural funding will not keep pace with increased demand.

Research Project Funding

- Substantial number of data centres reported that project funding was significant.
- What are projects for?
 - R&D: developing tools, systems and processes that can then be implemented in core service.
 - Business intelligence.
 - Staff and business development: 'preparing for the future'.
- Does this matter? Or is it healthy?
 - One respondent indicated at 70-30 structural/project split might be optimal: *'The goal would be to have a funding model that consists of 70% structural funding, and 30% project based funding.'*

One of the alternative funding models, mentioned by 25% of the repositories, is charging for deposit

Questions:

- Does data deposit offer a transparent and scalable way to cover cost?
- Are fees a barrier to deposit data for individual researchers?
- Does it require a huge administration that will cost a lot?
- Do the deposit fees cover all costs or only part of it?

Funding options under consideration

37

Exploring alternatives?

- 65% yes
 - 25% no
 - 10% a little
-
- Sponsorships
 - Contracts for specific services offered (hosting, archiving, curation)
 - Expanding the number of affiliated institutions
 - Deposit fees
 - Funders making more money available (given priority for data)
 - Specific services for the commercial sector (mentioned by one)
 - More services for national memory institutes

Some Trends Wrt Future of Funding:

- There is friction between the perpetuity objectives of digital preservation and the timebound funding of repositories:
“The cost of long term preservation is now only covered for the first five years of preservation. More data and thus higher costs are expected in the near future. Demands and requirements will grow.”
- Where funding is now sufficient, people are concerned how to accommodate the ever increasing data deluge and the costs involved:
“Stakeholder and data volumes are growing rapidly and funding not following.”
- The priority on data sharing now provides a positive atmosphere for more funding: *“there is clearly a growth in the market to provide curation and repository services to sponsors.”*
- However, these factors are fashion-prone and may come to an end: *“Not sure how [Funding Agency] will deal with core infrastructure funding”*

Some Trends Wrt Future Models of Funding:

- One issue is finding the time (under current contract obligations) to explore new funding models: *“We are not allowed to do the analysis of the real business model cost, this is not a contracted deliverable! Don't have the funding to be able to investigate feasible business models.”*
- Another issue is the lack of insight into a reasonable fee structure/amount: *“Trying to figure out the right cost for a Terabyte of data, or 7 hours of curation”*

- Sustaining Domain Repositories for Digital Data: A White Paper, December 11, 2013, Prepared by Carol Ember (HRAF, Yale University) and Robert Hanisch (VAO, Space Telescope Science Institute)
http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf
- Towards Sustainable Stewardship of Digital Collections of Scientific Data, Robert R. Downs, Robert S. Chen, 2013
<http://www.gsdi.org/gsdiconf/gsdi13/papers/130.pdf>
- Databases fight funding cuts, Online tools are becoming ever more important to biology, but financial support is unstable. Monya Baker, 05 September 2012,
<http://www.nature.com/news/databases-fight-funding-cuts-1.11347>
- Who will pay for public access to research data?, Francine Berman and Vint Cerf, 09 August 2013
<http://www.sciencemag.org/content/341/6146/616.full.pdf?keytype=ref&siteid=sci&ijkey=.e2Ezowko%2FxF2>

Publishing Data Workflows

Chairs: Sunje Dallmeier-Tiessen, Fiona Murphy, Theodora Bloom



it barriers

RESEARCH DATA ALLIANCE



WORLD DATA SYSTEM

Background and Motivation

- Only a small fraction of research data is preserved and shared, often with a bare minimum of metadata
- Often due to the lack of “established” or “trusted” services and workflows

But there are established or emerging workflows!

- Usually in selected disciplines, e.g., Earth Sciences
- Some provide credit via citation mechanisms

Objectives

- Provide an analysis of a representative range of existing and emerging workflows and standards for data publishing
 - Including deposit and citation
 - Provide reference models, a “classification”
- Test implementations of key components for application in new workflows
- Illustrate the benefits of the reference models for researchers and organisations

Relevance

- Information about workflows crucial for researchers and other stakeholders to understand the options available to practice open science
- Helps to illustrate different possibilities for data sharing, leading to more efficient and reliable reuse of research data
- Shows those involved in research data where they fit in the overall scheme of things

More detailed work programme

- Identification of a smaller set of reference models covering a range of such workflows to include:
 - For example, **when and where** QA/QC and data peer-review fit into the publishing process
 - **Who** does what and when...
 - Automated vs. “manual” processes
- Selection of key use cases and organizations in which components of a reference model can be **implemented** and tested for suitability
 - For example: dedicated data peer review
 - For example: metadata checks

First results of workflow analysis

File Edit View Insert Format Data Tools Add-ons Help Last edit was made 6 days ago by anonymous

| A | B | C | D | E | F | G | H | I | J |
|--|---|---|---|---|---|--|--|--|--|
| Some notes beforehand: | | | | | | | | | |
| - please feel free to add workflows you think should be part of the analysis by expanding the table to the right - if you think, there are categories missing (or need to be changed), please add a row and drop an email to the list to highlight the change that affects the other ongoing analysis - but please c - please provide as many links to documentations, charts as possible to enrich the analysis | | | | | | | | | |
| | Workflow name | STFC Data centre | NSIDC Data centre | ENVRI reference model Infrastructure/S ervice provider (data centre/repositor y?) | OJS/ Dataverse Data repository | INSPIRE Digital library | NPG (PubChem & Scientific Data) Publisher | UK Data Archive/Service provider/ data centre | PREPARDE (NCAR Cisl) Data centres (+ data journals) |
| | Stakeholder in charge [inst, publisher, disciplinary repository, ...] | | | | | disciplinary, international data repository | | | |
| | Available online | http://www.prelida | ftp://sidads.colora | http://confluence.e | http://projects.iq.harvard.edu/ojs-dyn&https://docs.google.com/document/d/1T-i2a4synXlhe3DCiYyALi8VYgh2hLdJJMmd6KVXhc/edit#&https://docs.google.com/file/d/0BzeLxEN77UZoYnBLWXpodmpTLTQ | inspirehep.net | http://www.dcc.ac | ODIN Deliverable [link?] | http://www.le.ac.uk/projects/preparde http://proj.badc.rl.ac.uk/preparde/attachment/wiki/DeliverablesList/D2_1_D2_2_PREPARDE_Workflows_combined_draft1.pdf |

Workflows in the current list

- STFC Data centre
- NSIDC Data centre
- ENVRI reference model
- OJS/ Dataverse
- INSPIRE Digital library
- NPG (PubChem & Scientific Data) Publisher
- UK Data Archive/Service
- PREPARDE (NCAR CISL)
- Ocean Data Publication Cookbook (UNESCO IOC)
- PURR Institutional repository
- ICPSR
- Edinburgh Datashare
- F1000 Research
- Ubiquity Press: Open Health Data Journal+...
- PANGAEA - Data Publisher for Earth and Environmental Sciences
- WDC Climate - Data Publisher for Climate Sciences
- CMIP / IPCC DDC - International project series in Climate Sciences
- GigaScience
- Dryad digital repository with integrated journals workflow
- Stanford Digital Repository
- Academic Commons: Columbia University Institutional Research Repository
- Elsevier: Data in Brief
- Integrated data publishing solution at Elsevier [through “traditional” journals]

Categories we are looking at

- Discipline
- Function of workflow
- PID assignment to dataset
- PID type -- e.g., DOI, ARK, etc.
- Peer review of data (e.g., by researcher & editorial review)
- Curatorial review of metadata (e.g., by institutional or subject repository?)
- Technical review & checks (e.g., for data integrity at repository/data centre on ingest)
- Discoverability: Indexing of the data -- if yes, where?
- Formats covered
- Persons/Roles involved, e.g., editor, publisher, data repository manager, etc.
- Link to data paper or “standalone” data
- Links to grants, usage of author PIDs
- Data citation facilitated
- Data life cycle referred to
- Standards compliance

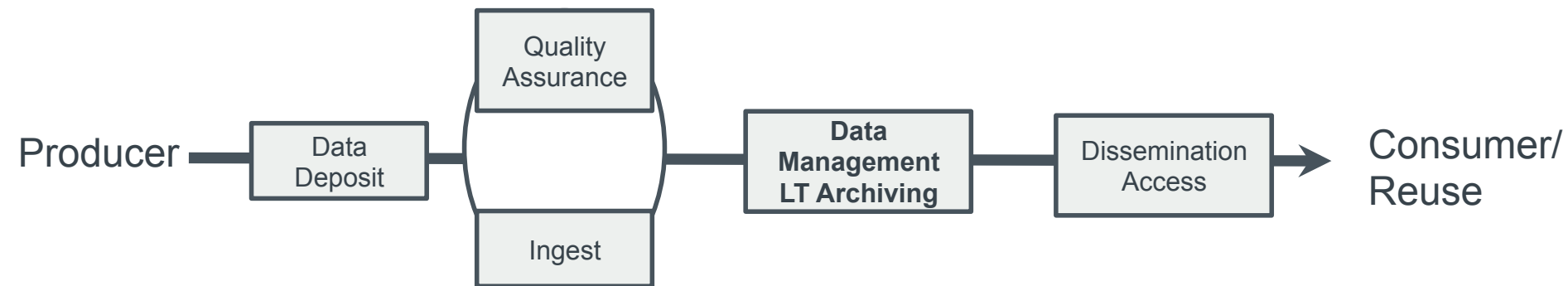
Observations

- The researcher/author generally initiates the workflow
- Discipline-specific repositories have the most rigorous ingest and review processes -- more general institutional repositories have a lighter touch
- Journals vs. repositories: For the former, any peer review is conducted externally, for many of the latter it is internal

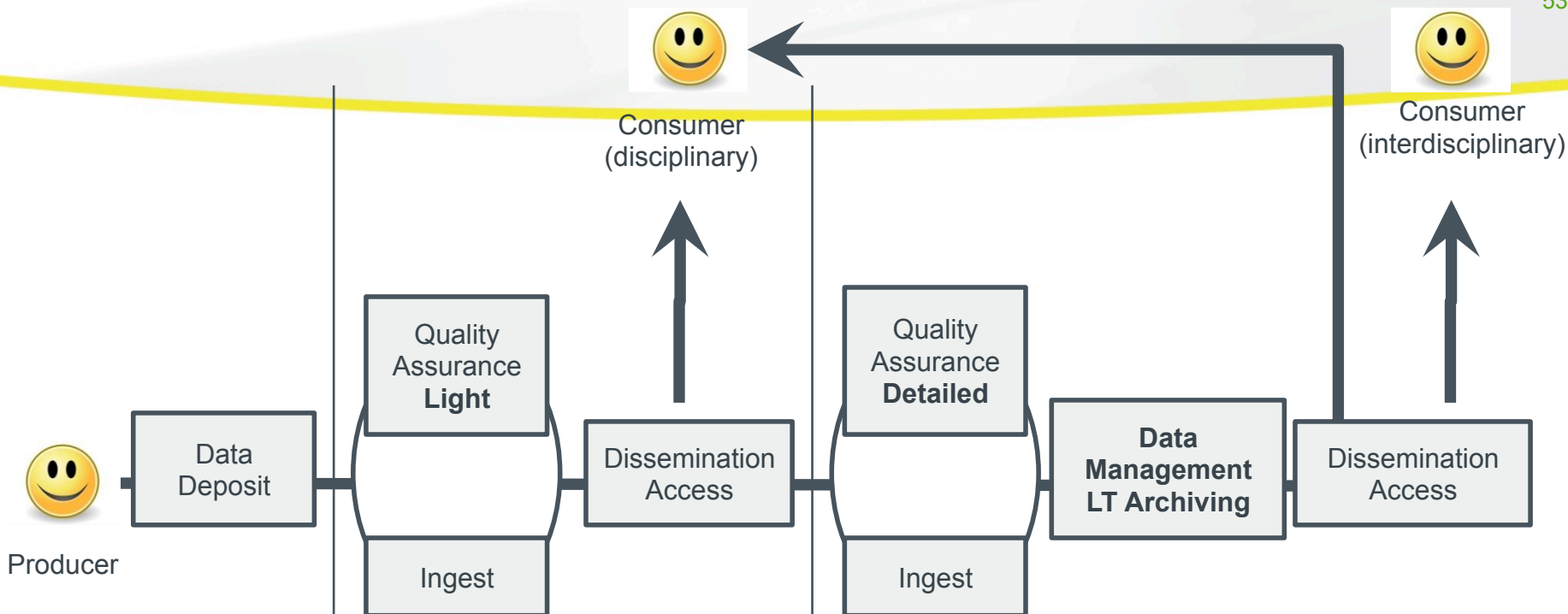
Repository

Simplified generic repository workflow

Researcher with a central role: submission/deposition



Review/QA mainly internal



Project Repositories:

- Data **are** published in a federated data infrastructure
- Data **are added and corrected**
- **Poor documentation**
- **Usually no data backup**
- Light-weight quality assurance against intl. and project standards
- Tendency that the project data never **become** stable
- Currently **no PIDs** assigned or reserved but **Handles planned**

Long-term Archive:

- Data **are** archived for the long term at a single location
- Data **are stable and curated**
- **Detailed documentation**
- **Data backup/redundancy**
- Quality assurance process is more detailed and includes a review
- Data is a “snapshot” of the project data at a certain time
- **DOIs** assigned to data collections

Designed by
M. Stockhausen

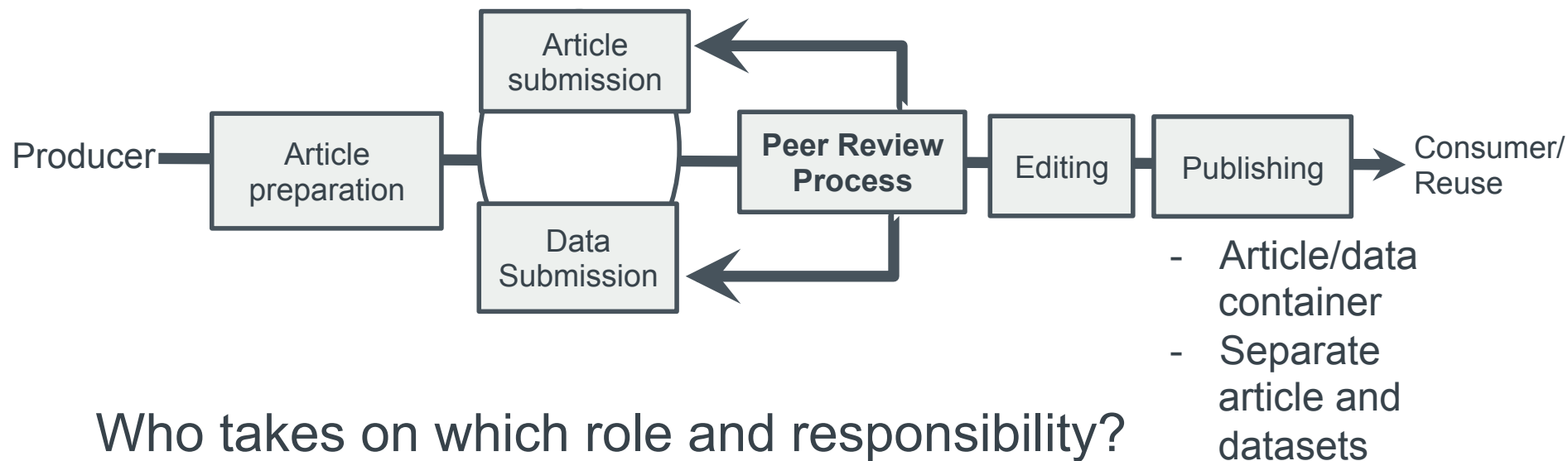
Lessons Learnt and questions

- Very diverse landscape
- Discipline-specific and cross-discipline actions
- Quality assurance a big topic in discipline-specific repositories
- Widespread persistent identification
- Data citation awareness
- Challenge: Bidirectional data-publication linking
- Challenge: Versioning

Publisher's perspective

Simplified generic publisher workflow

Researcher takes over several roles: submitter, reviewer, editor potentially?



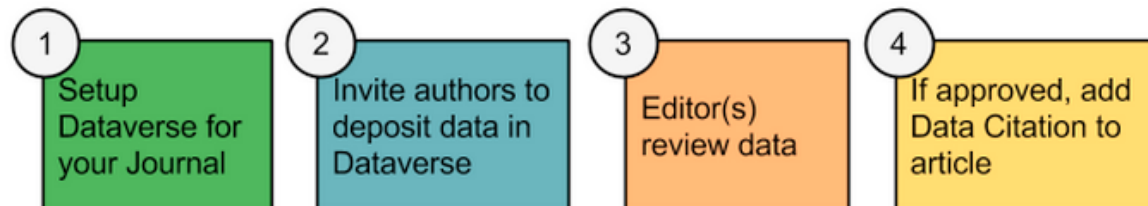
Who takes on which role and responsibility?

Example Workflows in Dataverse: Connect Data to Journals

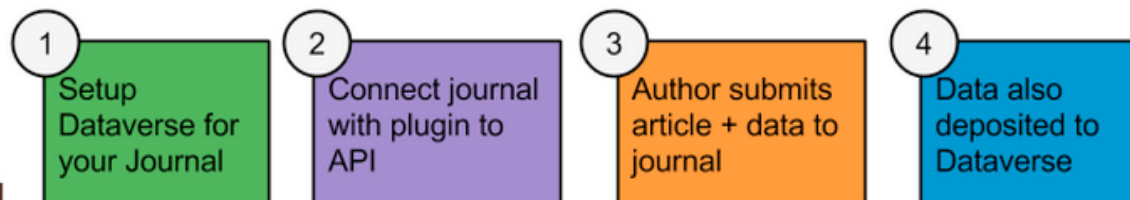
A. Journals include Dataverse as a Recommended Repository



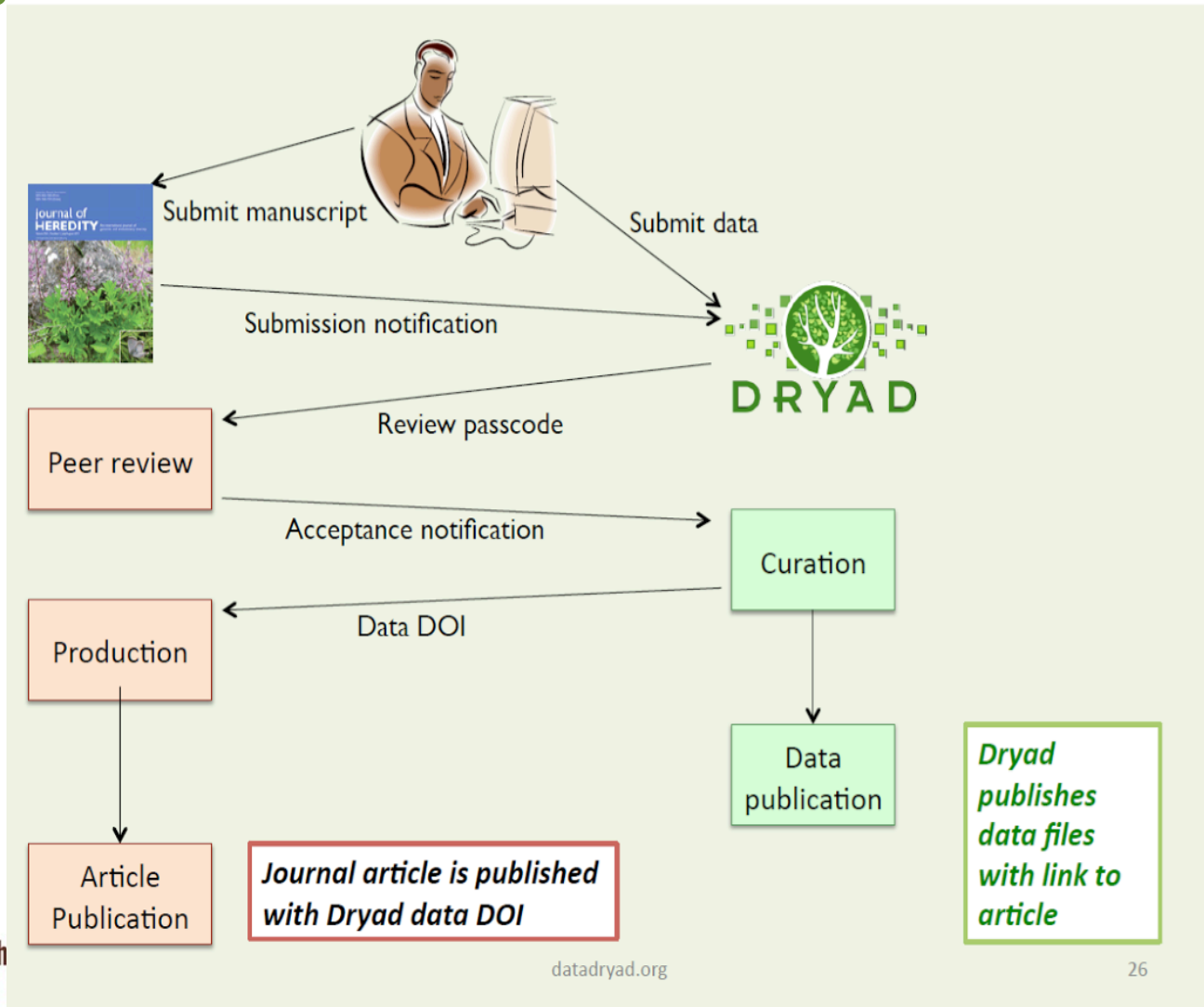
B. Authors Contribute Directly to a Journal's Dataverse



C. Automated Integration of Journal + Dataverse (e.g., OJS)



Example: Dryad repository integrated with journals



Lessons learnt and questions

- Recommended repositories for collaboration? Who decides/how?
- External review
 - Open, plus invitation
 - Closed, upon invitation
 - Blind
- Emerging data and software journal landscape: no information yet on uptake

Current and future work

How to get involved

- Contribute to the workflow analysis: <http://bit.ly/1BBQQPW>
- Contribute your own workflow “walk-throughs” and use cases
- Tell us what is needed for a “successful” workflow in your institute/ discipline

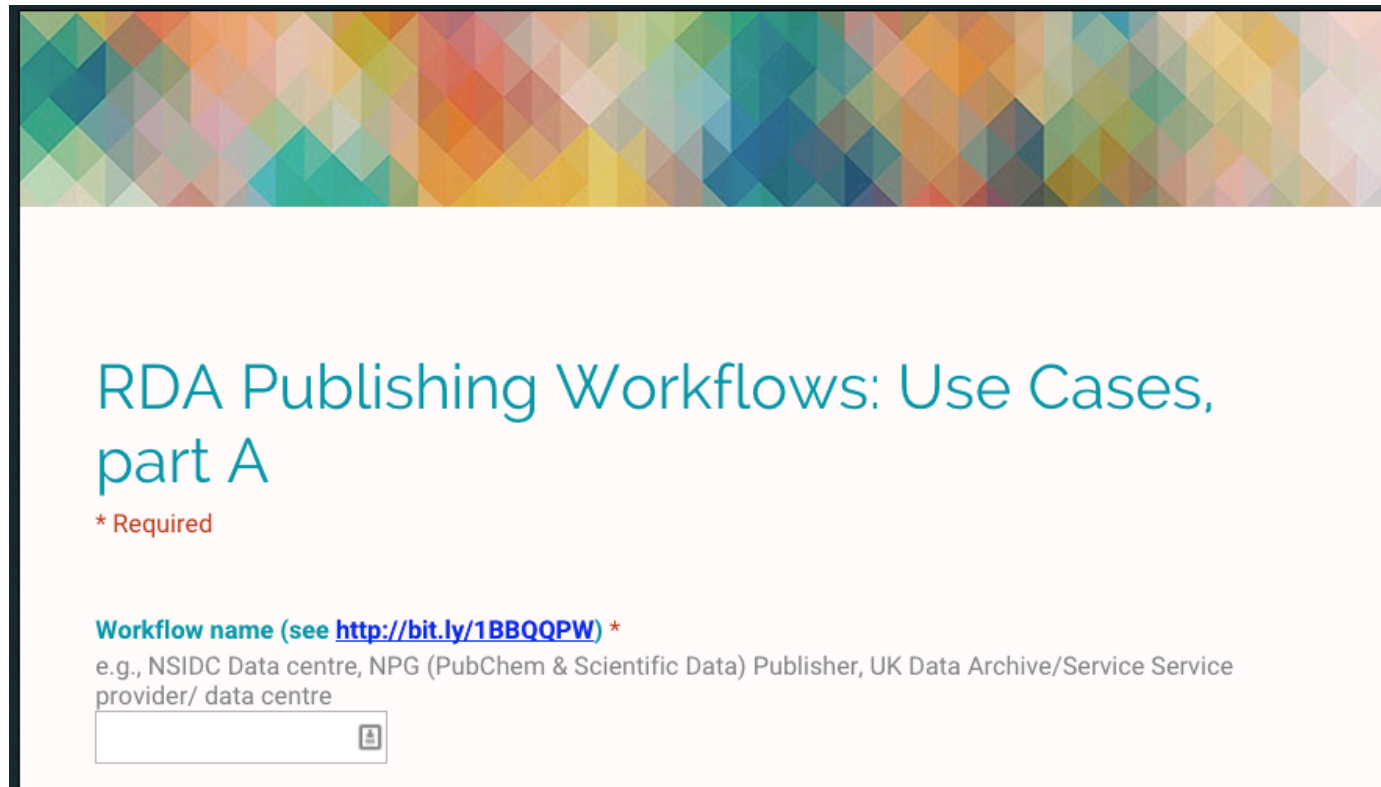
... Moving to implementation

- Tell us if you are interested to learn from a specific example or are maybe considering implementing data publishing workflows
- Tell us if you have code/documentation to share

Use case development

The tools: Part A

<http://goo.gl/forms/Wkc7KyxvX5>



RDA Publishing Workflows: Use Cases,
part A

* Required

Workflow name (see <http://bit.ly/1BBQQPW>) *

e.g., NSIDC Data centre, NPG (PubChem & Scientific Data) Publisher, UK Data Archive/Service Service provider/ data centre

The tools: Part A

<http://goo.gl/forms/Wkc7KyxvX5>

Primary actor (role) *

- ☐ Data consumer
- ☐ Data producer
- ☐ Funder
- ☐ Institution
- ☐ Librarian
- ☐ Publisher
- ☐ Repository
- ☐ Reviewer
- ☐ Other:

The tools: Part A

<http://goo.gl/forms/Wkc7KyxvX5>

Case description *

Please provide a brief description of the scenario envisioned for this use case.

Trigger *

Please describe the condition that causes the use case to start.

Pre-condition *

Please describe what must be in place before the use case can be started, or what constraints must be met before initiation.

Post-condition *

Please describe the final desired outcome(s) that indicate(s) the use case has been successfully completed, or what constraints must be met by use case execution.

The tools: Part A

<http://goo.gl/forms/Wkc7KyxvX5>

List of actors *

Please identify all of the actors participating in this use case.

- ☐ Data consumer
- ☐ Data producer
- ☐ Funder
- ☐ Institution
- ☐ Librarian
- ☐ Publisher
- ☐ Repository
- ☐ Reviewer
- ☐ Other:

The tools: Part A

<http://goo.gl/forms/Wkc7KyxvX5>

Thank you! You have completed Part A of this use case. For the next part, you will be completing multiples of a form, to address each individual actor listed in this use case. Click this to get to Part B: <http://goo.gl/forms/ZFRrzG6krX>

The tools: Part B

<http://goo.gl/forms/ZFRrzG6krX>



RDA Publishing Workflows: Use Cases, part B

Greetings, and welcome to part B! You've completed a parent case in part A, and will create several children for it in part B by filling out a part B form for each of the actors listed in the use case detailed in part A.

The tools: Part B <http://goo.gl/forms/ZFRrzG6krX>

Case ID # *

This is a bit complicated. Go to the link and get the Case ID # from the parent case you completed in part A.

https://docs.google.com/spreadsheets/d/1RaWINBw822Z7ra6oLfZ6IMZA0_r9ms8FIhFyLQipCCE/edit?usp=sharing

The tools: Part B <http://goo.gl/forms/ZFRrzG6krX>

Actor *

Please enter the actor name (role) for this use case scenario.. This is a value from the previous form (part A) item, "List of Actors". Each part B should have a different actor. Each actor from part A should have an associated part B.

Needs *

What does this actor need to accomplish?

Goals *

Why does this actor need to accomplish those things? What are the motivations?

Stakeholders *

Who are other stakeholders in this process from this actor's perspective?

Priority *

Considering the actors in this use case, what priority should this actor be given?

Extensions *

Are there any points where the behavior of this actor might influence or extend to affect the behavior of another actor? (If not, please answer 'No'.)

The tools: Part B <http://goo.gl/forms/ZFRrzG6krX>

Repository Features necessary for this actor *

Check which repository features are necessary for this actor in this case. Please add features that are not listed.

- ☐ PID assignment to data set
- ☐ Peer review of data, (e.g. by researcher & editorial review)
- ☐ Curatorial review of metadata (e.g. by institutional or subject repository?)
- ☐ Technical review & checks (e.g. for data integrity at repository/data centre on ingest)
- ☐ External indexing of the data
- ☐ Links to paper
- ☐ Links to grants
- ☐ Links to other related resources
- ☐ Usage of author PIDs
- ☐ Data citation facilitated
- ☐ Standards, such as OAIS
- ☐ Other:

Trusted data publishing contains:

72

- Standardized information about the data
 - Disciplinary standards
 - Basic common metadata sets
- Distinct Roles, Workflows and Responsibilities
 - Authorship, Submission
 - Curation
 - Quality Assurance
 - Peer review
- Persistent Identification
 - Permanent reference
 - Data citation

Final steps: towards a reference description for data publication

- Finalize terminology and clarify connotations
- What is used today/tomorrow
- What works, What is recommended practice
- Role of individuals, repositories and publishers
- Recommendations/Conclusions



Bibliometrics WG

Chairs: Sarah Callaghan
(sarah.callaghan@stfc.ac.uk)

Kerstin Lehnert
(lehnert@ldeo.columbia.edu)

Todd Carpenter
(tcarpenter@niso.org)

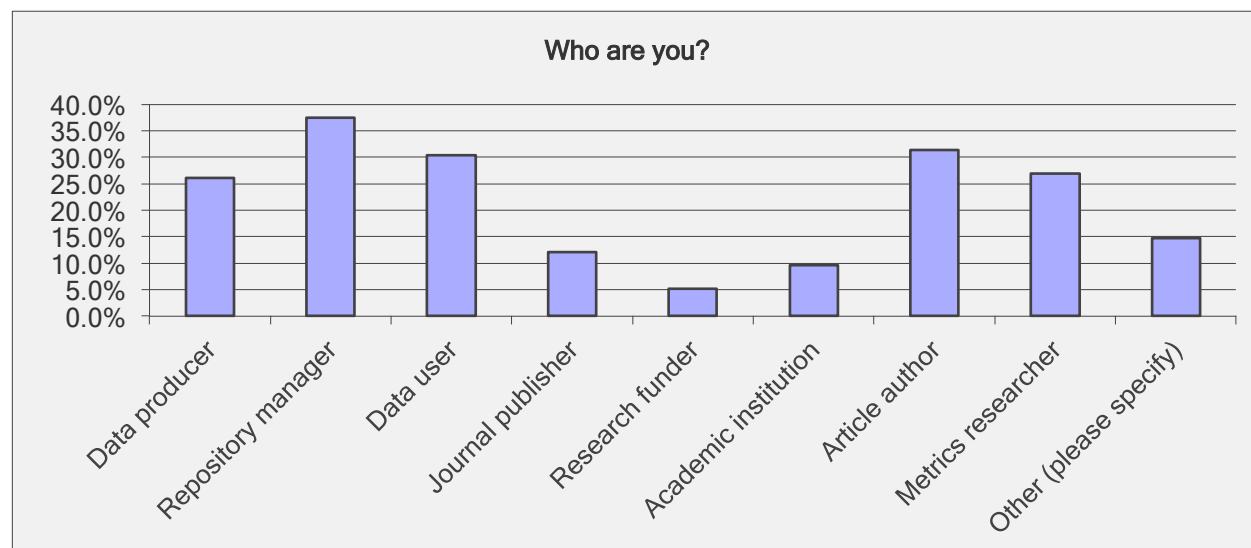
John Kratz
(johnkratzcdl@gmail.com)

- Quantitative measures for the use, utility, and impact of data.
- **Rationale:** Essential for a culture change toward full appreciation and recognition of data as a part of the scholarly record.
 - raise the value of data acquisition, curation, and sharing;
 - encourage more and better data citation;
 - augment the overall availability and quality of research data;



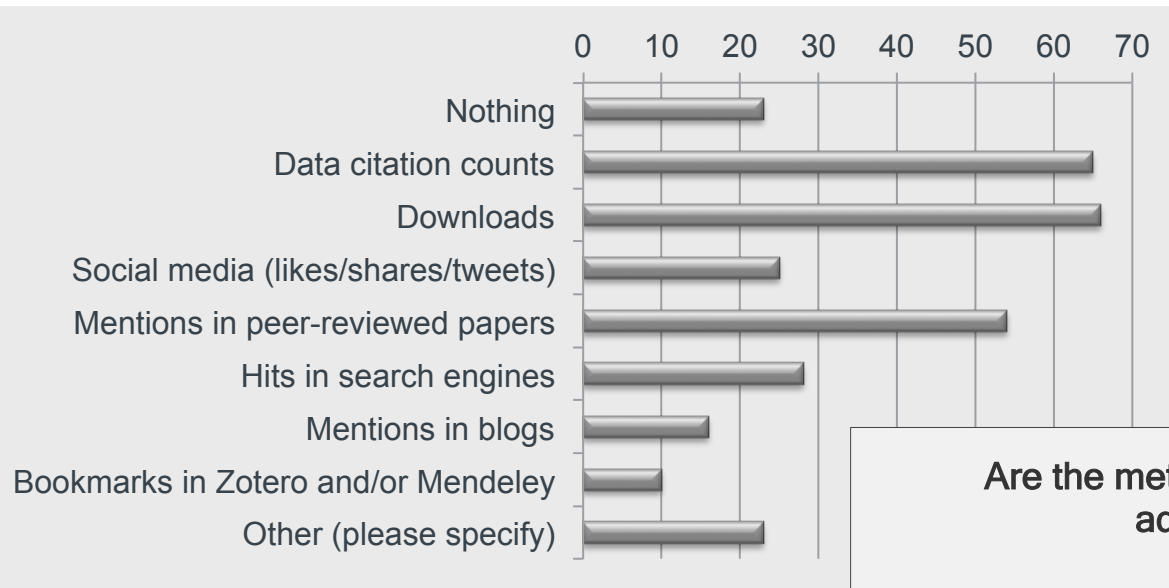
- **“conceptualize data metrics and corresponding services to overcome barriers”**
 - Summarize current and emerging *data citation practices* of all stakeholders (journals, data centers, funders, societies);
 - Understand and articulate *necessary organizational and cultural changes* in the scholarly publishing system needed to foster proper attribution of data sets;
 - Evaluate and report on *possible models* how data use and citations can be successfully tracked and measured based on existing and emerging approaches (e.g. altmetrics);
 - Identify and report on *possible barriers* for the implementation and adoption of data citation and data metrics solutions.

- 42 members of WG
- survey completed by 115 participants
 - presented in webinars, at FORCE 2015, at RDA4 & 5



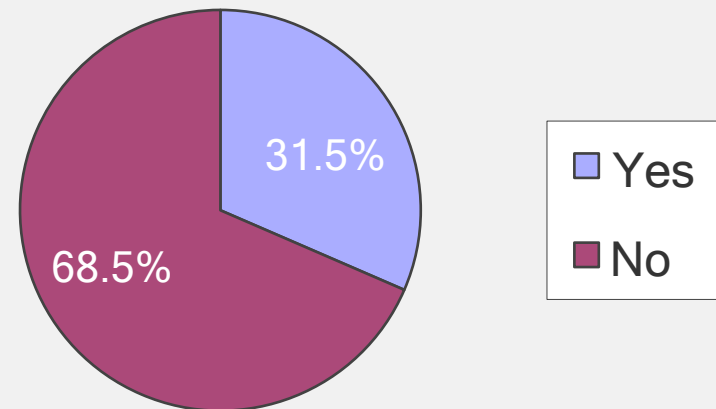
Survey Results: Mostly Expected

78



- **Citations** are preferred metrics, downloads next.
- **Standards** are missing.
- **Culture change** is needed.

Are the methods you use to evaluate impact adequate for your needs?



- Partner with DataCite to study resolution data of DataCite DOIs.
 - Prepare paper on how much data resolution activity is taking place.
 - Lay out the start of longitudinal study on data use/citation using DOI resolution data.
 - In discussion with DataCite about getting raw resolution data (from stats.datacite.org). Danielle Pollock (PhD student, University of Tennessee), working on this from July.
 - Repository practices may influence DOI resolutions.

- Establish a partnership with the PLOS “Making Data Count” project <http://articlemetrics.github.io/MDC/> on metrics of data sets. NISO project on Altmetrics. Casrai Data Level Metrics study.
 - Recruit RDA members to trial the ‘Making Data Count’ open source tools (scheduled to be released in Spring 2015)
 - Get the tools validated and adopted by repositories – what’s needed to be put into the tool for wider adoption? Will a one size fits all technical solution work given the diversity of the repositories?
 - Project: **Giving Researchers Credit for their Data (JISC Research Data Spring)** – project with similar interests. Survey coming out soon – WG members respond to it!
 - Begin to gather/analyze data generated by these services (lay out groundwork of a longitudinal study on data use/reuse based on data collected).
 - Take existing PLOS DLM tool and adapt to datasets. Use DataONE and Dryad as proof of concept. dml.plos.org – link to tool in progress
- PLOS survey been written up, but not published yet – hopefully out soon. Not that much overlap with our WG survey

- Continue to monitor adoption of bibliometrics for data.
 - Survey/desk research the repository community on the types of usage data they have, they could analyze, or could share.
 - Collect data from different repositories
 - Need standard set of questions What do repositories collect, and what might they be able to collect?
 - Connect with RDA IG on domain repositories
- Translate our insights into recommendations

Thank you!

82

Any questions?

[sarah.callaghan](mailto:sarah.callaghan@stfc.ac.uk)
[@stfc.ac.uk](mailto:sarah.callaghan@stfc.ac.uk)

