How PID are DOI’s?

Brian Kirkegaard Lunn,
Senior Metadata Manager, Digital Science
PID Apalooza 2019, Dublin
Focus of this session

- Analysis of 1.7 million DOI’s originating from the community
- Focus on the invalid DOI’s among these 1.7 million
- Highlight knowledge about the quality of this well adapted PID when adapted to local systems
  - How big a percentage of DOI’s are invalid?
  - Is there a pattern of these invalid DOI’s
Data integrated in multiple dimensions

Pre-publication: 1-5 years from grant to publication

Post-publication:
- immediate
- 2-3 years
- years
- years
- decades

- NIH Grants
- Research Conferences
- Data sets Publications
- Tweets Blogs
- Clinical Trials
- Citations
- Patents
- Policy docs

Dimensions
The data and links driving Dimensions...

- **98M Publications**
  - Improved metadata of 65m
  - 12M links
  - 1.0bn links
  - 700K links
  - 399K links

- **4.1M grants**
  - $1.3 trillion in funding
  - 16M to funders
  - 370K to funders

- **37M patents**
  - 318M links
  - 11M links
  - 10M links

- **384K Policy papers**
  - 202k to funders links
  - 865k links

- **84M Altmetric mentions**
  - 9.9M links
  - 1.0bn links

- **426K Clinical trials**
  - 155K links

- **98M Publications**
  - Improved metadata of 65m
  - 1.0bn links
  - 700K links
  - 399K links

- **37M patents**
  - 318M links
  - 11M links
  - 10M links

- **384K Policy papers**
  - 202k to funders links
  - 865k links

- **84M Altmetric mentions**
  - 9.9M links
  - 1.0bn links

- **426K Clinical trials**
  - 155K links
The Publications of Dimensions...

The publications data spine is based on
- Crossref
- PubMed

Enhanced by processing full text where available
- OA publications
- Non-OA publications made available by publishers for the purpose

90 million publications with a DOI

98M Publications improved metadata of 65m
1.0bn links
Data for the analysis - background

Coverage analysis for institutions

- Free analysis of coverage in Dimensions of any institutions list of publications
- Requires a list of DOI from the institution in question
- This list is then matched against publications in Dimensions
- Result of analysis provided back to institution
Data for analysis - overview

51 cases of coverage analysis picked for a representative sample
- Across types of institutions (Archive, Education, Facility, Government, Healthcare)
- Across regions (The Americas, Europe, Asia, Australia)

Selection also based on
- DOI’s in the list - 34k doi’s on average per institution (ranging from 48 to 150k)
- Coverage - 97.3% doi’s covered in Dimensions on average (ranging from 86.5% to 99.6%)
  - ~50% of uncovered DOI’s are invalid (based on ~100 per institution, only)
  - List of doi’s delivered by institutions comprise on average 45% of publications associated with the institution in question (ranging from 1% to 88%)
Essentially what is the data corpus?

DOI’s as they are available in local publications systems (CRIS, Excel, …)
- As they are added and used by institutions
- Including copy/paste and import from external systems
- Covering 1.7 million DOI’s in total - 1.6 million unique DOI’S

Data corpus
- 26,290 unique DOI’s not covered in Dimensions found in the 51 coverage analyses
  - Excluding DOI’s with a DOI alias covered in Dimensions
  - 1.6% of unique DOI’s
26k non-covered doi’s
(1.6% of the 1.6 million unique DOI’s analysed)

Valid DOI but not covered in Dimensions
- 12k - 46%

Invalid doi’s
- 825 - 3%

DOI does not exist
- 13k - 51%
12k Valid DOI’s not found in Dimensions
(46% of the 26k non-covered DOI’s analysed - 0.75% of unique DOI’s)
825 Invalid DOI’s
(3% of the 26k non-covered DOI’s analysed)

DOI’s that do not follow the DOI format with a prefix and a suffix
- ~80% can be converted to DOI format with intellectual intervention
  ○ Missing 10.
  ○ Missing / as delimiter between prefix and suffix
  ○ Missing . as delimiter in prefix
- ~20% have nothing to do with DOI’s
  ○ Text
  ○ URL
  ○ ISBN / ISSN
  ○ Number
13k DOI’s does not exist
(51% of the 26k non-covered DOI’s analysed)

- DOI’s do not resolve to any publication
- DOI’s are valid with respect to format
- ~5k DOI contain reserved or unsafe characters (e.g., ?, #, [, ] and white space)
  - 10.1002/(sici)1097-0215(19960220)69:1<64::aid-ijc17>3.3.co;2-#
  - 10.1002/(sici)1097-0215(19960220)69:1%3C64::aid-ijc17%3E3.3.co;2-%23
Large part of DOI’s seem to originate from publishers

- 511 sici DOI - publications with prefix/(sici)[issn]([date])
  - 23 sici that are valid DOI’s
- Example 10.1002/(sici)1096-8628(19960122)

Count of non-existing DOI’s per prefix

- Elsevier (10.1016), 1817 (14%)
- Wiley (10.1002), 1075 (8.3%)
- Springer (10.1007), 711 (5%)
- American Meteorological Society (10.1175), 678 (5%)
- Wolters Kluwer Health, (10.1097), 505 (4%)
- Taylor & Francis, (10.1080), 479 (3%)

13k DOI’s does not exist
(51% of the 26k non-covered DOI’s analysed)
Ghost publications?

PID’s are really persistent

Errors in DOI
- 10.17088/tksyms.2016..45.004
- 10.17088/tksyms.2016..45.004

DOI’s per publication format -
Take aways

- 99% of DOI from analysed research organisations are valid
- Only a very small fraction of DOI’s across the 51 institutional DOI lists are invalid
  - 14k => 0.9% of the 1.6m doi’s analysed
- Only a very small fraction (~0.05%) seem to be invalid due to human manipulation
- A large portion of the invalid DOI’s appear to be correctly formatted DOI’s
  - Following the DOI format but not recognised as an existing DOI
- Is there a need for cleaning institutional held DOI’s?
  - a DOI laundromat
Questions/Comments?