



PIDs and Preservation

Incorporating persistent identifiers in a preservation strategy

Moderator: Craig van Dyck

Panellists: Amanda Bartell, Helena Cousijn, Ivo Wijnbergen

PIDS and Preservation



Crossref and preservation

Amanda Bartell, Head of Member Experience, Crossref
iPres, September 2019



Mission

Crossref makes research outputs easy to find, cite, link, assess, and reuse.

We're a not-for-profit membership organization that exists to make scholarly communications better.

Open for Science

BROWSE

GATEWAYS & COLLECTIONS

HOW TO PUBLISH

ABOUT

BLOG

MY RESEARCH

SIGN IN

Check for updates

RESEARCH ARTICLE

REVISED

Predicting transcription factor binding using ensemble random forest models [version 2; peer review: 2 approved with reservations]

✉ Fatemeh Behjati Ardakani

✉ Florian Schmidt

✉ Marcel H. Schulz

1-3*

1-4*

1,2,5

* Equal contributors

✉ Author details

This article is included in the **RPackage** gateway.

This article is included in the **Max Planck Society** collection.

METRICS

471

VIEWS

44

DOWNLOADS

Get PDF

Get XML

Cite

Export

Track

Email

Open Peer Review

Reviewer Status

?

?

1

Reviewer Reports

	Invited Reviewers	
	1	2
Version 2 (revision) 02 Sep 19		
Version 1 04 Oct 18	<div>?</div> <div>read</div>	<div>?</div> <div>read</div>

1. Jan Grau

Martin Luther University of Halle-Wittenberg (MLU), Halle, Germany

2. Garu D. Storms

Washington University

How to cite: Behjati Ardakani F, Schmidt F and Schulz MH. Predicting transcription factor binding using ensemble random forest models [version 2; peer review: 2 approved with reservations]. *F1000Research* 2019, 7:1603 (<https://doi.org/10.12688/f1000research.16200.2>)

Furthermore, we show that incorporating DNase1-seq peaks is essential to reduce the false positive rate of TF binding predictions compared to considering the raw DNase1 signal.

Conclusions: Analysis of important features reveals that the models preferentially select motifs of other TFs that are close interaction partners in existing protein-protein-interaction networks. Code generated in the scope of this project is available on GitHub: <https://github.com/SchulzLab/TFAnalysis> (DOI: 10.5281/zenodo.1409697).

Keywords

ENCODE-DREAM in vivo Transcription Factor binding site prediction challenge, Transcription Factors, Chromatin accessibility, Ensemble learning, Indirect-binding, TF-complexes, DNase1-seq

✉ Corresponding authors: Fatemeh Behjati Ardakani, Florian Schmidt, Marcel H. Schulz

Competing interests:

No competing interests were disclosed.

Grant information:

This work was supported by the Cluster of Excellence on Multimodal Computing and Interaction (DFG) [EXC248]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

©

Copyright: © 2019 Behjati Ardakani F et al. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite:

Behjati Ardakani F, Schmidt F and Schulz MH. Predicting transcription factor binding using ensemble random forest models [version 2; peer review: 2 approved with reservations]. *F1000Research* 2019, 7:1603 (<https://doi.org/10.12688/f1000research.16200.2>)

Sign up for content alerts

Browse by related subjects

Bioinformatics

Chromatin

DNA sequencing

Transcription factors



PID myths

- Myth 1: DOI = Preservation
- Myth 2: DOI = automatic, magical persistence



Layers of persistence

- The identifier
- The content
- The mechanism for updating the mapping between the content and the identifier
- The metadata about the content



Content: Defunct DOI page

DOI Not Found

10.5555/jioj

This DOI cannot be found in the DOI System. Possible reasons are:

- The DOI is incorrect in your source. Search for the item by name, title, or other metadata using a search engine.
 - The DOI was copied incorrectly. Check to see that the string includes all the characters before and after the slash and no sentence punctuation marks.
 - The DOI has not been activated yet. Please try again later, and report the problem if the error continues.
-

You may report this error to the responsible DOI Registration Agency using the form below. Include your email address to receive confirmation and feedback.

DOI:

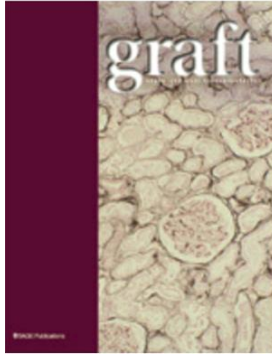
URL of Web Page Listing the DOI:

Your Email Address:

Additional Information About the Error:



Content: Third party archive solution



Graft: Organ and Cell Transplantation is no longer in publication. Online access to this title is now provided through the archiving organizations listed below.

Determining Significant Physiologic Incompatibilities

Hammer, Claus et al.

Graft(2001),4(2):108

<http://dx.doi.org/10.1177/152216280100400206>

[| Free Article - EDINA |](#)

[| Article Available - Portico |](#)

[| Free Article - Stanford University |](#)

[| Article Available - Koninklijke Bibliotheek |](#)



```
1 // 20190816102004
2 // http://api.crossref.org/works?filter=has-archive:true&facet=publisher-name:*&rows=0
3
4 {
5   "status": "ok",
6   "message-type": "work-list",
7   "message-version": "1.0.0",
8   "message": {
9     "facets": {
10       "publisher-name": {
11         "value-count": 252,
12         "values": {
13           "Wiley": 2016157,
14           "S. Karger AG": 216972,
15           "American Geophysical Union (AGU)": 104079,
16           "John Benjamins Publishing Company": 33950,
17           "eLife Sciences Publications, Ltd": 7110,
18           "PeerJ": 6305,
19           "Australasian Medical Publishing Co. Pty Ltd. (AMPCo)": 6256,
20           "AIP Publishing": 5565,
21           "EMBO": 4725,
22           "Beilstein Institut": 4628,
23           "Oxford University Press (OUP)": 2669,
24           "American Society for Microbiology": 2096,
25           "Finnish Society of Forest Science": 2017,
26           "Hikari, Ltd.": 1121,
```

"total-results": 2427817,



Mechanism for updating the mapping between the content and the DOI

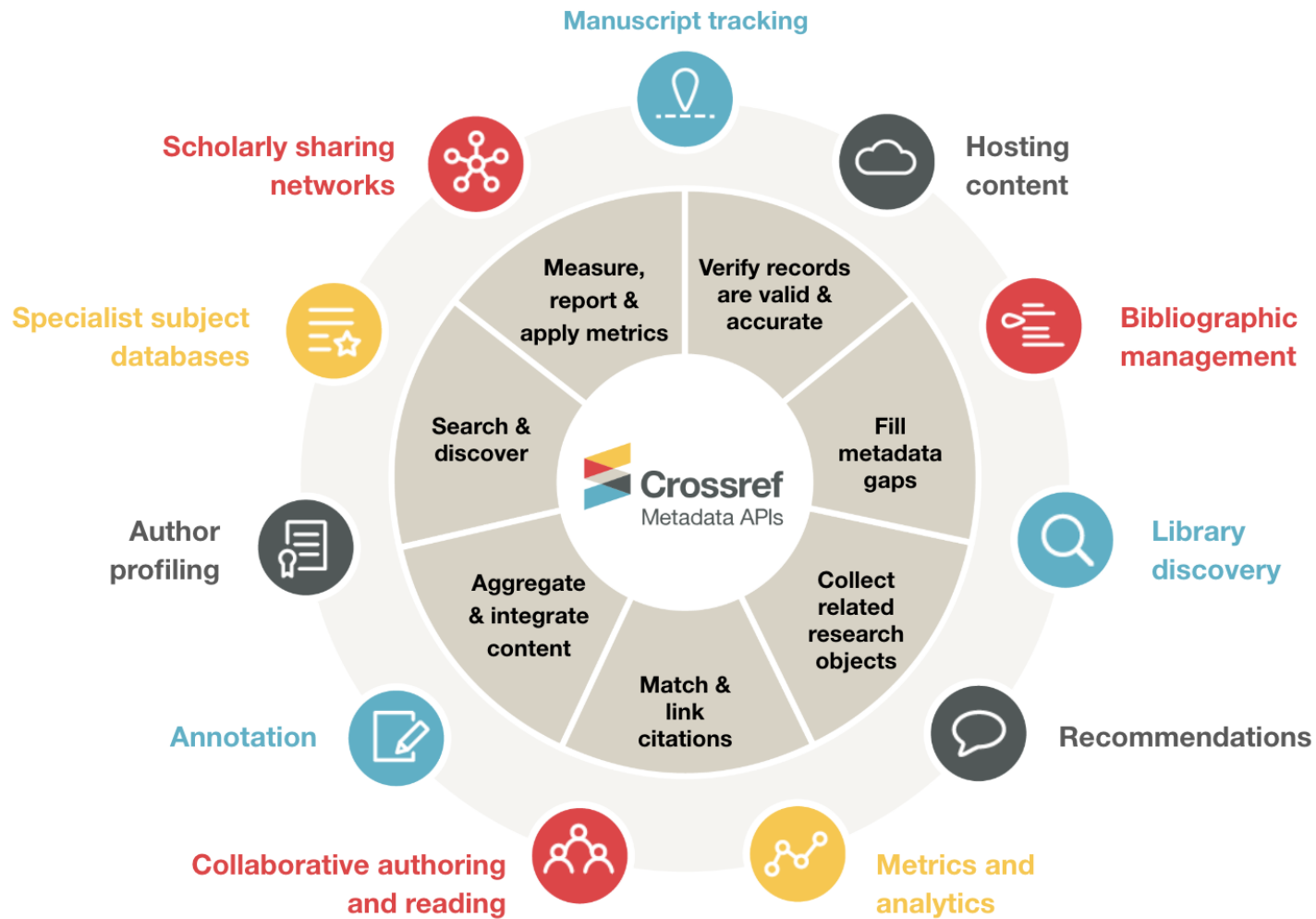
- That would be us
 - Open
 - Not-for-profit
 - Community-governed
 - Community-led
 - Not grant funded



Metadata about the content



<https://www.crossref.org/members/prep/>



@abartell



DataCite & Preservation

September 19,
2019

About DataCite



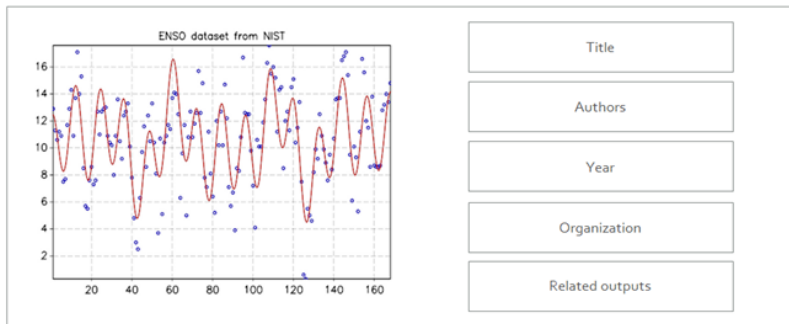
Connecting research, Identifying knowledge.

DataCite's mission is to be the world's leading provider of persistent identifiers for research. Through our portfolio of services, we provide the means to create, find, cite, connect, and use research. We seek to create value and develop community-driven, innovative, open, integrated, usable, and sustainable services for research.

The workflow



1. Take a scholarly resource



2. Describe it

3. Assign a DOI



4. DataCite Search Index

ATLAS Collaboration, "Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC: $H \rightarrow \gamma\gamma$,"
<http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44>

5. Enjoy the benefits

Google Dataset Search

Indexing Services

Track citations

Gather Usage Stats



Unique



Persistent

PIDs & FAIR



(meta)data are assigned a globally unique and persistent identifier



(meta)data are retrievable via an identifier using a standardized protocol
metadata are accessible, even when the data are no longer available



(meta)data use a formal, accessible, shared, and broadly applicable language knowledge representation.
(meta)data include qualified references to other (meta)data



meta(data) are described with a plurality of accurate & relevant attributes
(meta)data are associated with detailed provenance

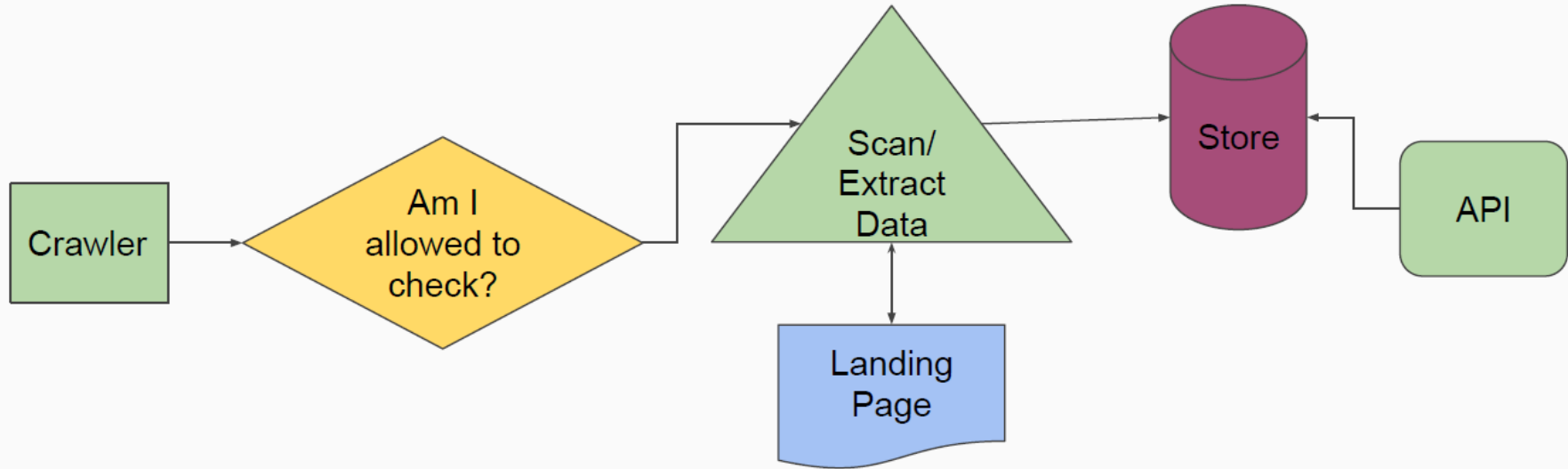
Preservation and access



> **preservation** aimed at providing **access** over time

> **access** depends upon **preservation** at a point in time

Link checking



Link checking



Link Checker

The URL [https://pure.strath.ac.uk/portal/en/datasets/edge-and-spatiallydependent-emission-optical-spectra-for-ingan-microleds-integrated-ultrathin-colloidal-quantum-dot-functionalised-glass\(28b1e9db-91f4-4787-9ef3-522401895239\).html](https://pure.strath.ac.uk/portal/en/datasets/edge-and-spatiallydependent-emission-optical-spectra-for-ingan-microleds-integrated-ultrathin-colloidal-quantum-dot-functionalised-glass(28b1e9db-91f4-4787-9ef3-522401895239).html) registered for this DOI was last checked **a month ago** on August 10, 2019.

✔ The URL resolves properly.

The link check redirected 2 times ([https://pure.strath.ac.uk/portal/en/datasets/edge-and-spatiallydependent-emission-optical-spectra-for-ingan-microleds-integrated-ultrathin-colloidal-quantum-dot-functionalised-glass\(28b1e9db-91f4-4787-9ef3-522401895239\).html](https://pure.strath.ac.uk/portal/en/datasets/edge-and-spatiallydependent-emission-optical-spectra-for-ingan-microleds-integrated-ultrathin-colloidal-quantum-dot-functionalised-glass(28b1e9db-91f4-4787-9ef3-522401895239).html), <https://pureportal.strath.ac.uk/en/datasets/28b1e9db-91f4-4787-9ef3-522401895239>), and resolved to <https://pureportal.strath.ac.uk/en/datasets/edge-and-spatially-dependent-emission-optical-spectra-for-ingan-m> with HTTP status code 200.

✔ The URL resolves to a landing page.

The link check returned the HTTP content type **text/html**.

✔ The landing page includes the DOI.

The link check found the DOI in the landing page HTML.

○ The landing page includes metadata in schema.org format.

The link check found no embedded JSON-LD with @context "<http://schema.org>". Please reach out to DataCite Support if we missed embedded schema.org metadata.

Community



ORCID

Connecting Research
and Researchers

Tackling Information Overload

ORCID is a non-profit organization, which provides a fully open and interoperable identifier to reliably connect researchers with their research contributions. The ORCID iD is a 16-digit identifier that researchers can register for and use for free.

Connects individuals and their professional contributions across disciplines, organizations, and time



Enables recognition of all types of research contributions and innovation

Helps research institutions, funders, publishers, and other organizations better track and support research work

How ORCID Works



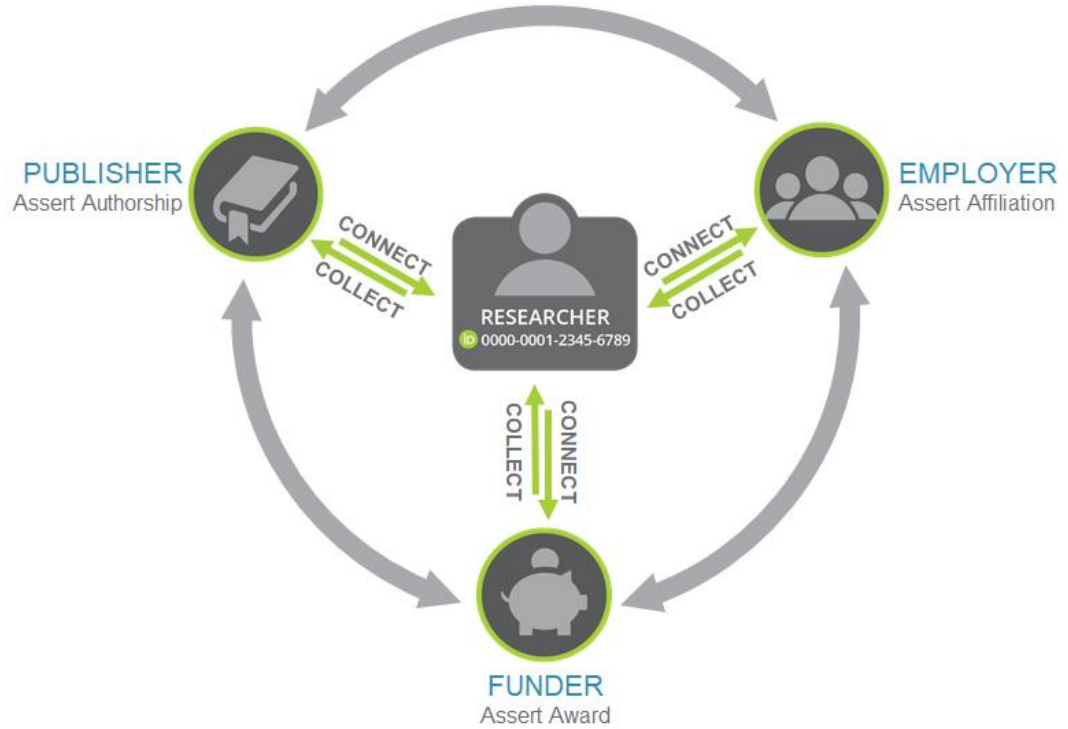
- It's a **registry** of unique persistent identifiers for researchers
- It's a **hub** that connects researchers with their professional activities and contributions
- It's a global **community** that enables researchers to share their data with other individuals, organizations, and systems

Interoperability

Sharing of information about research across systems

People - Places - Things

- People
→ ORCID
- Places
→ ROR
Research Organization Registry
A community-led project to develop an open, sustainable, usable, and unique identifier for every research organization in the world. (ror.org)
- Things
→ Crossref & DataCite



ORCID and Digital Preservation

Our challenge:

We are a claims-repository, not a content-repository. We preserve ORCID metadata, not the location the metadata points to.

Main problems we think about:

- What is our responsibility?
- Is the metadata enough?
- Is there a need to ensure preservation of the actual content we point to?
And how can we help?

A group of hands holding up large, red, three-dimensional block letters that spell out the phrase "WE WANT YOU". The letters are arranged in a slightly staggered fashion, with some hands holding multiple letters. The background is plain white.

WE WANT YOU

To help us find answers to these problems