Free your papers, researchers!

Dissemin team (Ryan Lahfa)

July 22, 2016

Introduction: what is a researcher? A Pokemon? Not yet.

Research is:

> *The systematic investigation into and study of materials and sources in order to establish facts and reach new conclusions.*

– *Oxford Dictionnaries*, `http://www.oxforddictionaries.com/definition/english/research`

What is exactly a paper ?

Who was there at the keynote on the gravitational waves?

What is exactly a paper ?

Who was there at the keynote on the gravitational waves?

This is a **breakthrough**, and there was a paper published for it!

Let's take a look to the full text here: `https://v.gd/7o6YaS`

What do we do with papers?

- We read papers to inform ourselves on what is going on in the field.

What do we do with papers?

- We read papers to inform ourselves on what is going on in the field.
- We cite papers in our thesis, in our bibliography.

What do we do with papers?

- We read papers to inform ourselves on what is going on in the field.
- We cite papers in our thesis, in our bibliography.
- We even build software using papers! (machine learning, database systems for example)

What do we do with papers?

- We read papers to inform ourselves on what is going on in the field.
- We cite papers in our thesis, in our bibliography.
- We even build software using papers! (machine learning, database systems for example)

What do we do with papers?

- We read papers to inform ourselves on what is going on in the field.
- We cite papers in our thesis, in our bibliography.
- We even build software using papers! (machine learning, database systems for example)

There is a catch, though. Research financed from public money is sometimes published through companies (Elsevier) or organizations (ACM, IEEE).

What do we do with papers?

- We read papers to inform ourselves on what is going on in the field.
- We cite papers in our thesis, in our bibliography.
- We even build software using papers! (machine learning, database systems for example)

There is a catch, though. Research financed from public money is sometimes published through companies (Elsevier) or organizations (ACM, IEEE).

These publishers decide to keep the papers behind a paywall, as if it was “closed-source”. So that these papers are not accessible, nor open.

Why Open Access is necessary?

Open Access is a really *important* concept for research:

- students can access those papers **because** their school pays for subscriptions to these publishers.

What about others? They simply cannot access or have to pay a ridiculous amount ($30 for 10 pages!) to access a PDF file (which was financed through public money).

Guess game! (students, you don't play.)

Publishers edit journals, and accessing to their content requires a subscription.

The most famous is Nature, published by Nature Publishing Group.

Guess game! (students, you don't play.)

Publishers edit journals, and accessing to their content requires a subscription.

The most famous is Nature, published by Nature Publishing Group.

So, in your opinion, how much does a subscription cost *per year* for **one journal** ?

Guess game! (students, you don't play.)

Publishers edit journals, and accessing to their content requires a subscription.

The most famous is Nature, published by Nature Publishing Group.

So, in your opinion, how much does a subscription cost *per year* for **one journal** ?

Well, over \$10 000 per year, you can have also journals peaking at over \$25 000.

[1] 2007–2008 data

Guess game! (students, you don't play.)

Publishers edit journals, and accessing to their content requires a subscription.

The most famous is Nature, published by Nature Publishing Group.

So, in your opinion, how much does a subscription cost *per year* for **one journal** ?

Well, over \$10 000 per year, you can have also journals peaking at over \$25 000.

According to Right to Research, Elsevier (a publisher) has around 31.7 %[1] of profit margin.

[1] 2007–2008 data

Guess game! (students, you don't play.)

Publishers edit journals, and accessing to their content requires a subscription.

The most famous is Nature, published by Nature Publishing Group.

So, in your opinion, how much does a subscription cost *per year* for **one journal** ?

Well, over \$10 000 per year, you can have also journals peaking at over \$25 000.

According to Right to Research, Elsevier (a publisher) has around 31.7 %[1] of profit margin.

What was Google's approximate profit margin in 2008 ?

[1] 2007–2008 data

Guess game! (students, you don't play.)

Publishers edit journals, and accessing to their content requires a subscription.

The most famous is Nature, published by Nature Publishing Group.

So, in your opinion, how much does a subscription cost *per year* for **one journal** ?

Well, over \$10 000 per year, you can have also journals peaking at over \$25 000.

According to Right to Research, Elsevier (a publisher) has around 31.7 %[1] of profit margin.

What was Google's approximate profit margin in 2008 ? 30.6 %.

[1] 2007–2008 data

Summary

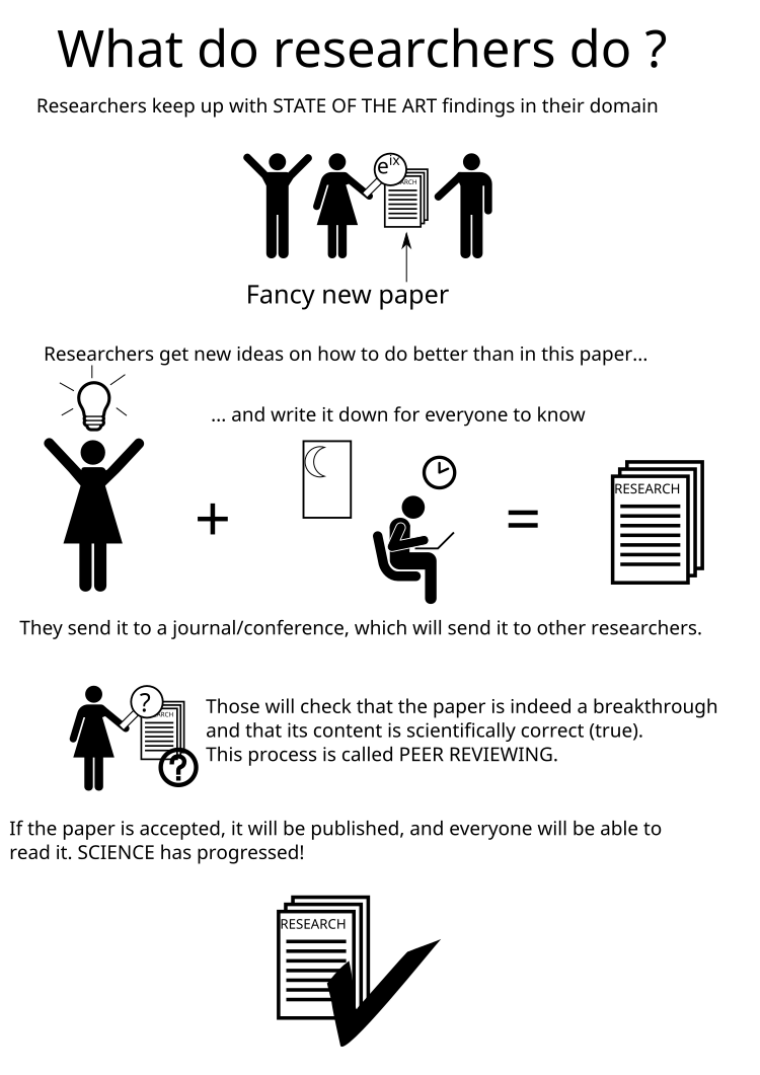


Figure 1: What we believe

Summary

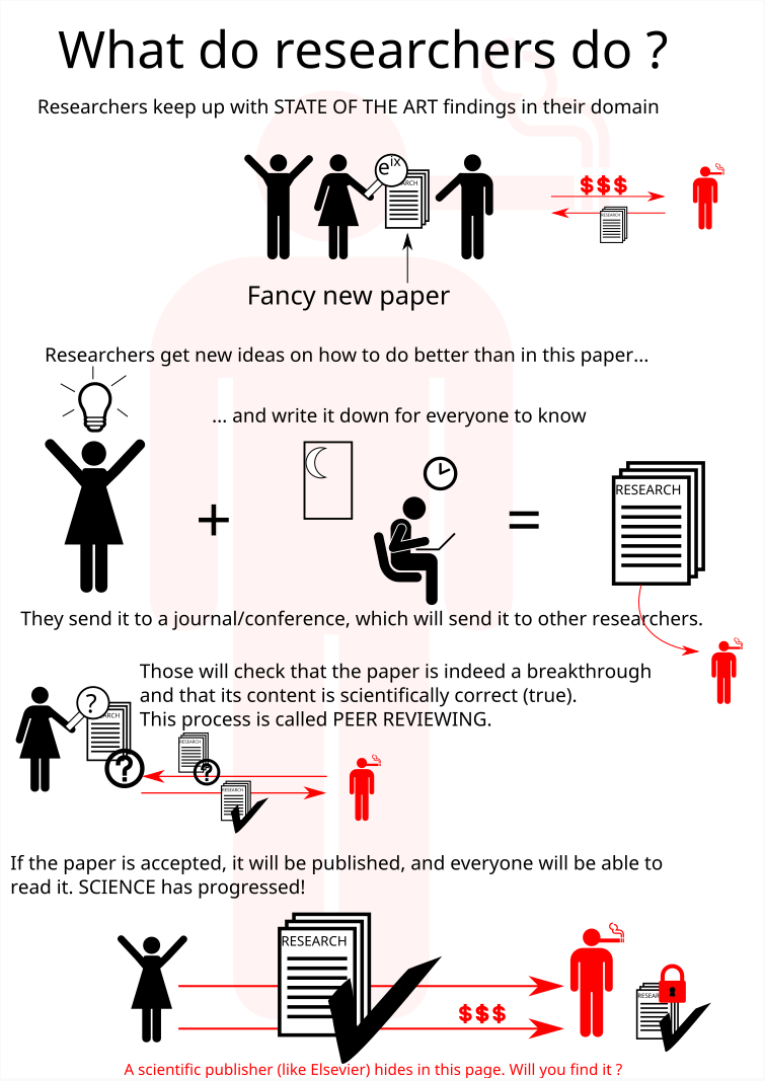


Figure 2: What we see

What are the consequences?

- Subscriptions are extremely expensive, even though the papers have been given away for free, so that researchers perform peer review on them.

What are the consequences?

- Subscriptions are extremely expensive, even though the papers have been given away for free, so that researchers perform peer review on them.
- Why shouldn't students from developing countries have access crucial papers which are behind a paywall?

What are the consequences?

- Subscriptions are extremely expensive, even though the papers have been given away for free, so that researchers perform peer review on them.
- Why shouldn't students from developing countries have access crucial papers which are behind a paywall?
- Why cannot people who aren't students access papers?

What are the consequences?

- Subscriptions are extremely expensive, even though the papers have been given away for free, so that researchers perform peer review on them.
- Why shouldn't students from developing countries have access crucial papers which are behind a paywall?
- Why cannot people who aren't students access papers?
- This could be you. As a developer, you can run into a situation where you need a paper and it is not available, only behind a expensive paywall!

What can we do to improve Open Access?

Using open source software,

Dissemin (`http://dissem.in/`) is a tool to give control back to researchers. We would like to **promote** a global Open Access policy and **achieve it**.

What can we do to improve Open Access?

Using open source software,

Dissemin (`http://dissem.in/`) is a tool to give control back to researchers. We would like to **promote** a global Open Access policy and **achieve it**.

- We fetch your papers from different sources.

What can we do to improve Open Access?

Using open source software,

Dissemin (`http://dissem.in/`) is a tool to give control back to researchers. We would like to **promote** a global Open Access policy and **achieve it**.

- We fetch your papers from different sources.
- We check the policy on these papers.

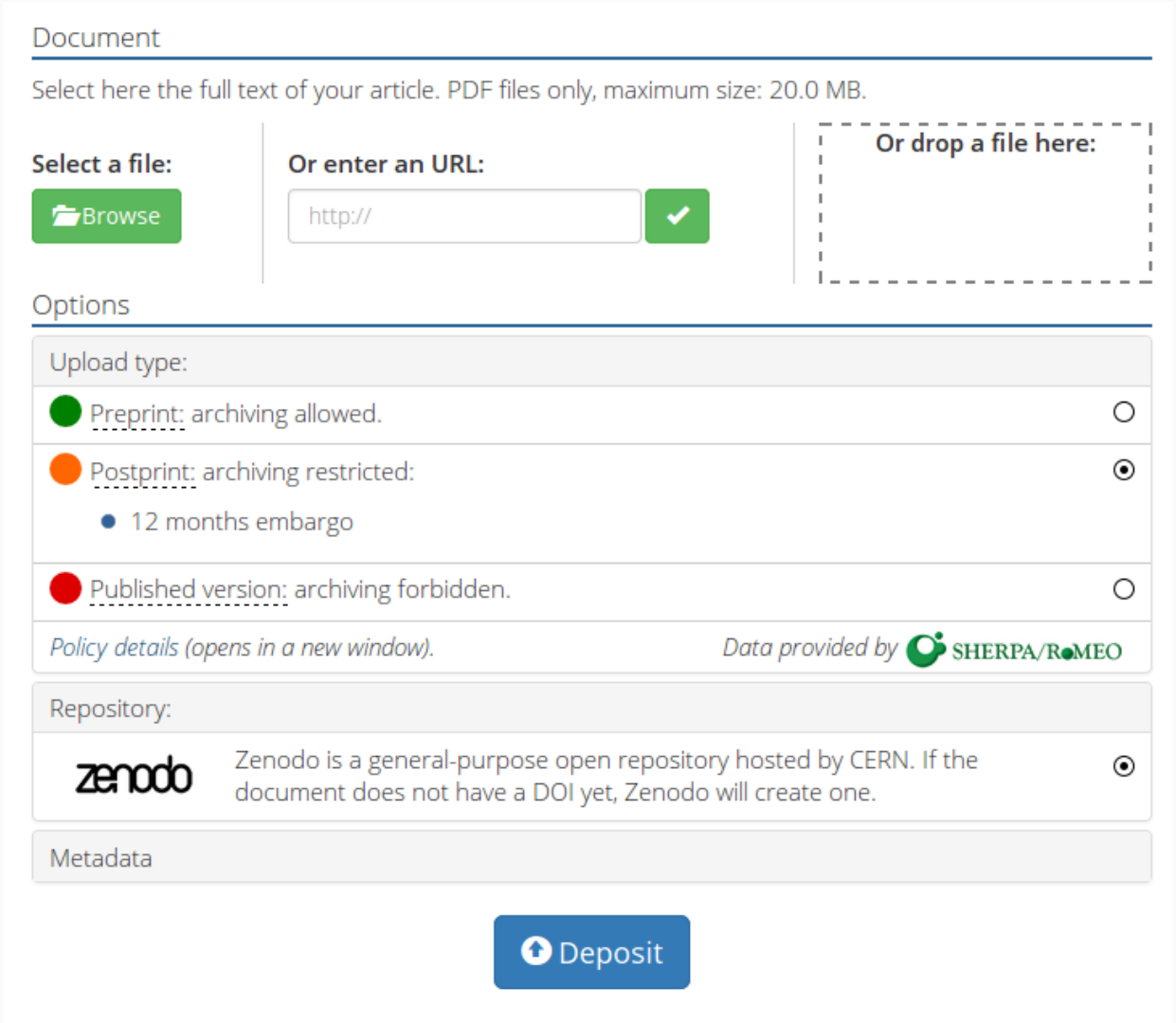
What can we do to improve Open Access?

Using open source software,

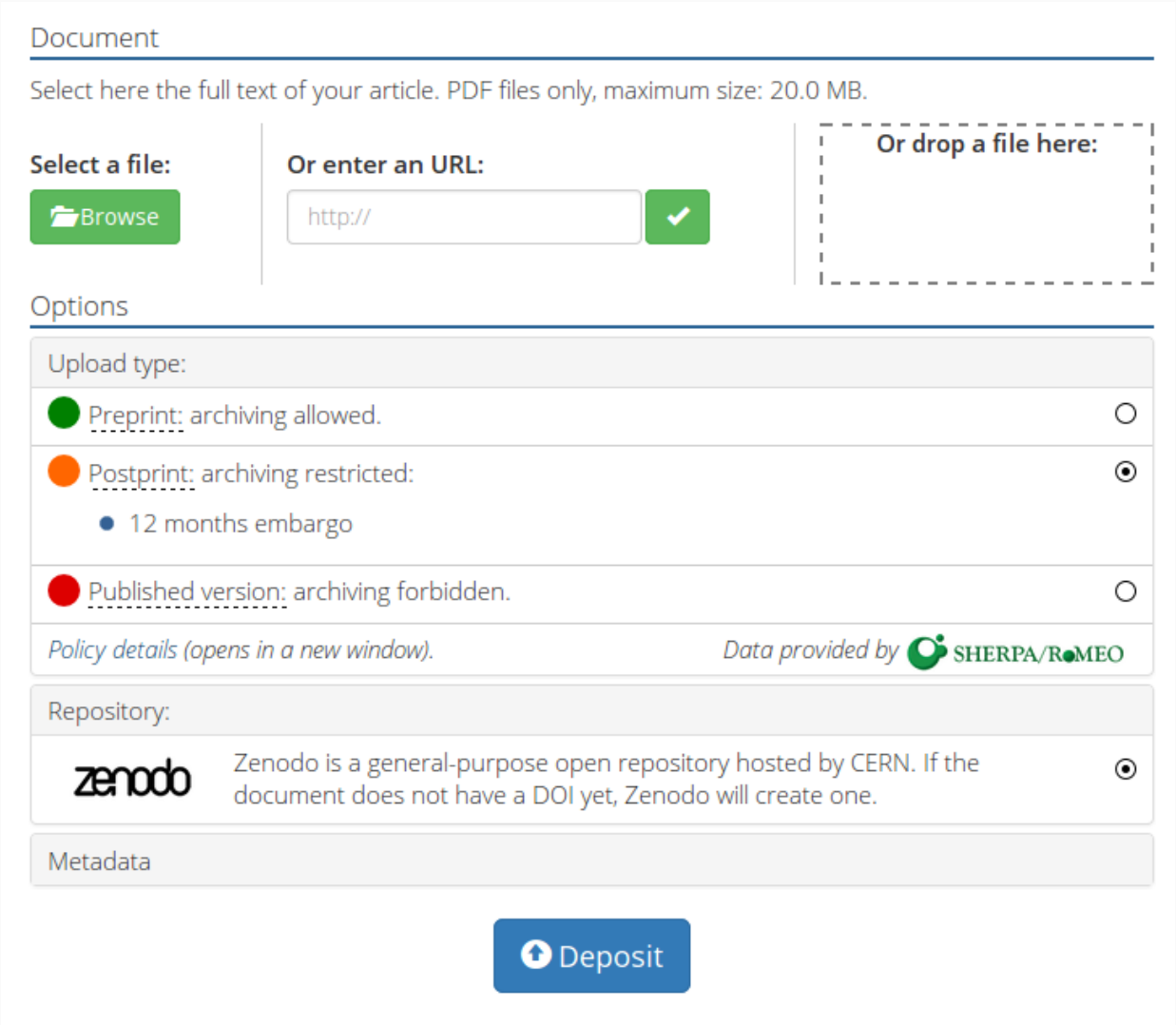
Dissemin (`http://dissem.in/`) is a tool to give control back to researchers. We would like to **promote** a global Open Access policy and **achieve it**.

- We fetch your papers from different sources.
- We check the policy on these papers.
- We tell you what you can deposit legally.

Upload!


Document
Select here the full text of your article. PDF files only, maximum size: 20.0 MB.
Select a file:
Browse
Or enter an URL:
http://
Or drop a file here:
Options
Upload type:
Preprint: archiving allowed.
Postprint: archiving restricted:
12 months embargo
Published version: archiving forbidden.
Policy details (opens in a new window).
Data provided by SHERPA/RoMEO
Repository:
zenodo
Zenodo is a general-purpose open repository hosted by CERN. If the document does not have a DOI yet, Zenodo will create one.
Metadata
Deposit

Upload!



Voilà! Your paper is free and accessible by everyone!

Who is behind Dissemin?

Dissemin is an initiative from a group of students of the "École Normale Supérieure" in France.

We are a non-profit organization participating in many Open Access related projects: Wikipedia, OpenCon, ...

This is a Python talk, where is Python?!

Dissemin is *of course* written in Python, using the Django framework!

We are using PostgreSQL to store papers and their metadata.

Challenge #1: Papers.

We have more than 15 millions metadata of papers and we are still getting more and more metadata through many academic sources.

But we have a problem.

Challenge #1: Papers.

We have more than 15 millions metadata of papers and we are still getting more and more metadata through many academic sources.

But we have a problem.

As you expect it, this amount of data is really non-trivial to handle, moreover metadata is more or less arbitrary in papers, so...

Challenge #1: Papers.

We have more than 15 millions metadata of papers and we are still getting more and more metadata through many academic sources.

But we have a problem.

As you expect it, this amount of data is really non-trivial to handle, moreover metadata is more or less arbitrary in papers, so...

We kept PostgreSQL and used its powerful JSON field!

PostgreSQL and JSON field

How do we use a JSON Field in Django?

PostgreSQL and JSON field

How do we use a JSON Field in Django?

```
class Paper(Model):

    authors_list = JSONField()
```

Awesome, you're done!

Amazing things about JSONField:

- Indexing works on JSON **sub**fields.

PostgreSQL and JSON field

How do we use a JSON Field in Django?

```
class Paper(Model):

    authors_list = JSONField()
```

Awesome, you're done!

Amazing things about JSONField:

- Indexing works on JSON **sub**fields.
- It's super efficient and can be your "NoSQL" world for a while!

PostgreSQL and JSON field

How do we use a JSON Field in Django?

```
class Paper(Model):

    authors_list = JSONField()
```

Awesome, you're done!

Amazing things about JSONField:

- Indexing works on JSON **sub**fields.
- It's super efficient and can be your "NoSQL" world for a while!
- Avoid very complex JOINS

PostgreSQL and JSON field

How do we use a JSON Field in Django?

```
class Paper(Model):

    authors_list = JSONField()
```

Awesome, you're done!

Amazing things about JSONField:

- Indexing works on JSON **sub**fields.
- It's super efficient and can be your "NoSQL" world for a while!
- Avoid very complex JOINS
- You can access subfields in queries directly without having to fetch the whole record!

Challenge #2 : Search have to be fast and relevant

With more than 15 millions of metadata, we have thought of many options, notably: PostgreSQL and its search engines (pg_trgm, full text search for example).

Challenge #2 : Search have to be fast and relevant

With more than 15 millions of metadata, we have thought of many options, notably: PostgreSQL and its search engines (pg_trgm, full text search for example).

This was not sufficient for the amount of data we had.

Challenge #2 : Search have to be fast and relevant

With more than 15 millions of metadata, we have thought of many options, notably: PostgreSQL and its search engines (pg_trgm, full text search for example).

This was not sufficient for the amount of data we had.

Enter Haystack.

Haystack and ElasticSearch

Haystack is a Python library which integrates with Django to provide awesome search tools.

Haystack and ElasticSearch

Haystack is a Python library which integrates with Django to provide awesome search tools.

- Multiple backends: ElasticSearch (the one we use), Solr, Whoosh, Xapian!

Haystack and ElasticSearch

Haystack is a Python library which integrates with Django to provide awesome search tools.

- Multiple backends: ElasticSearch (the one we use), Solr, Whoosh, Xapian!
- Faceting!

Haystack and ElasticSearch

Haystack is a Python library which integrates with Django to provide awesome search tools.

- Multiple backends: ElasticSearch (the one we use), Solr, Whoosh, Xapian!
- Faceting!
- Real-time indexing!

Haystack and ElasticSearch

Haystack is a Python library which integrates with Django to provide awesome search tools.

- Multiple backends: ElasticSearch (the one we use), Solr, Whoosh, Xapian!
- Faceting!
- Real-time indexing!

Haystack and ElasticSearch

Haystack is a Python library which integrates with Django to provide awesome search tools.

- Multiple backends: ElasticSearch (the one we use), Solr, Whoosh, Xapian!
- Faceting!
- Real-time indexing!

We are still working to make this faster and better, but we are really happy of the capabilities of these technologies!

Challenge #3 : PREVENT DUPLICATES PAPERS

A really hard feature is to prevent our database to be polluted with many duplicates due to the slightly variations in titles, partial authors lists, and a lot of things which makes the research world a lot funnier !

Challenge #3 : PREVENT DUPLICATES PAPERS

A really hard feature is to prevent our database to be polluted with many duplicates due to the slightly variations in titles, partial authors lists, and a lot of things which makes the research world a lot funnier !

- So we are using a fingerprinting technique, we have a function which takes a paper and reduce its minimal form (remove diacritics, lowercase, sort, simplify).

Challenge #3 : PREVENT DUPLICATES PAPERS

A really hard feature is to prevent our database to be polluted with many duplicates due to the slightly variations in titles, partial authors lists, and a lot of things which makes the research world a lot funnier !

- So we are using a fingerprinting technique, we have a function which takes a paper and reduce its minimal form (remove diacritics, lowercase, sort, simplify).
- And we compute a hash on it, here is our fingerprint!

Challenge #3 : PREVENT DUPLICATES PAPERS

A really hard feature is to prevent our database to be polluted with many duplicates due to the slightly variations in titles, partial authors lists, and a lot of things which makes the research world a lot funnier !

- So we are using a fingerprinting technique, we have a function which takes a paper and reduce its minimal form (remove diacritics, lowercase, sort, simplify).
- And we compute a hash on it, here is our fingerprint!
- Then, if we have a similar fingerprint in our database, we merge the paper!

Challenge #3 : PREVENT DUPLICATES PAPERS

A really hard feature is to prevent our database to be polluted with many duplicates due to the slightly variations in titles, partial authors lists, and a lot of things which makes the research world a lot funnier !

- So we are using a fingerprinting technique, we have a function which takes a paper and reduce its minimal form (remove diacritics, lowercase, sort, simplify).
- And we compute a hash on it, here is our fingerprint!
- Then, if we have a similar fingerprint in our database, we merge the paper!

Challenge #3 : PREVENT DUPLICATES PAPERS

A really hard feature is to prevent our database to be polluted with many duplicates due to the slightly variations in titles, partial authors lists, and a lot of things which makes the research world a lot funnier !

- So we are using a fingerprinting technique, we have a function which takes a paper and reduce its minimal form (remove diacritics, lowercase, sort, simplify).
- And we compute a hash on it, here is our fingerprint!
- Then, if we have a similar fingerprint in our database, we merge the paper!

So far, this technique is working *more or less* fine, we are always looking at how we can improve that. Especially when we have papers with very minimal metadata coming from some sources which makes our task harder!

Closing on challenges

We have many more challenges around machine learning to disambiguate name authors, perform title cleaning from LaTeX markup, infrastructure scripts (We have Vagrant for development, we would like Ansible for production), more deposit interfaces and sources!

Our GitHub repository is filled of interesting issues, we need your help : `https://github.com/dissemin/dissemin`

Closing on Open Access

We are a non-profit organization in France, having multiple projects around Dissemin :

- Proxy for DOI (Digital Object Identifier)

Closing on Open Access

We are a non-profit organization in France, having multiple projects around Dissemin :

- Proxy for DOI (Digital Object Identifier)
- Open Access bot for Wikipedia

Closing on Open Access

We are a non-profit organization in France, having multiple projects around Dissemin :

- Proxy for DOI (Digital Object Identifier)
- Open Access bot for Wikipedia
- Crawlers for repositories (Dublin Core for example)

Closing on Open Access

We are a non-profit organization in France, having multiple projects around Dissemin :

- Proxy for DOI (Digital Object Identifier)
- Open Access bot for Wikipedia
- Crawlers for repositories (Dublin Core for example)
- OAI-PMH protocol implementation

Inspiration (developers)

Like, Lasse Schuirmann from coala[2] Team which made a talk about "Growing an Open Source community" yesterday.

I want you to do something depending on what you prefer:

[2] coala must be written with a lowercase c, this is important really. Don't screw up.

Inspiration (developers)

Like, Lasse Schuirmann from coala[2] Team which made a talk about "Growing an Open Source community" yesterday.

I want you to do something depending on what you prefer:

If you are a developer interested in open access

- Clone Dissemin.

[2] coala must be written with a lowercase c, this is important really. Don't screw up.

Inspiration (developers)

Like, Lasse Schuirmann from coala[2] Team which made a talk about "Growing an Open Source community" yesterday.

I want you to do something depending on what you prefer:

If you are a developer interested in open access

- Clone Dissemin.
- Run it using Vagrant or anything.

[2] coala must be written with a lowercase c, this is important really. Don't screw up.

Inspiration (developers)

Like, Lasse Schuirmann from coala[2] Team which made a talk about "Growing an Open Source community" yesterday.

I want you to do something depending on what you prefer:

If you are a developer interested in open access

- Clone Dissemin.
- Run it using Vagrant or anything.
- Try it out and deposit fake papers for fun.

[2] coala must be written with a lowercase c, this is important really. Don't screw up.

Inspiration (developers)

Like, Lasse Schuirmann from coala[2] Team which made a talk about "Growing an Open Source community" yesterday.

I want you to do something depending on what you prefer:

If you are a developer interested in open access

- Clone Dissemin.
- Run it using Vagrant or anything.
- Try it out and deposit fake papers for fun.
- Take an issue and submit us a pull request.

[2] coala must be written with a lowercase c, this is important really. Don't screw up.

Inspiration (developers)

Like, Lasse Schuirmann from coala[2] Team which made a talk about "Growing an Open Source community" yesterday.

I want you to do something depending on what you prefer:

If you are a developer interested in open access

- Clone Dissemin.
- Run it using Vagrant or anything.
- Try it out and deposit fake papers for fun.
- Take an issue and submit us a pull request.
- If anything goes wrong, blame us and ping us.

[2] coala must be written with a lowercase c, this is important really. Don't screw up.

Inspiration (researcher)

If you are a researcher interested in open access

- Talk about Dissemin to everyone of your peers.

Inspiration (researcher)

If you are a researcher interested in open access

- Talk about Dissemin to everyone of your peers.
- Persuade them to open their papers.

Inspiration (researcher)

If you are a researcher interested in open access

- Talk about Dissemin to everyone of your peers.
- Persuade them to open their papers.
- Open your own papers.

Inspiration (researcher)

If you are a researcher interested in open access

- Talk about Dissemin to everyone of your peers.
- Persuade them to open their papers.
- Open your own papers.
- If anything goes wrong, complain to us!

Thank you EuroPython!

Contact us at team@dissem.in or @disseminOA