Validating and enriching VABB-SHW data through external sources

Raf Guns <raf.guns@uantwerpen.be>

ENRESSH training school, Poznań, 24 Oct 2019

Universiteit Antwerpen

ecoom

Validating enriching external sources

Manual

19. Complete missing data and validate the accuracy of metadata

"If the use of the actual research output is not feasible, complete missing metadata and validate the accuracy using other national and international bibliographic sources (e.g., CrossRef, WorldCat, Web of Science). Ideally, external data sources should be chosen on the basis of transparency and reliability of their data collection practice."

Manual

30. Follow and adapt to developments in research practices, research policy, and database maintenance

"Since the context of databases is subject to change, databases need to adapt to ensure that they can continue to fulfil their purposes."

Example

VABB-SHW: Comprehensive database of SSH publications, Flanders (Belgium)

- Created in context of Flemish funding model to cover SSH, as a complement to Web of Science
- 2019: funding system reform

  "The international collaboration parameter C3 calculates the share of each university in the number of publications of which one or more addresses on it […] belong to the university, together with at least one foreign address […]"

Exercise

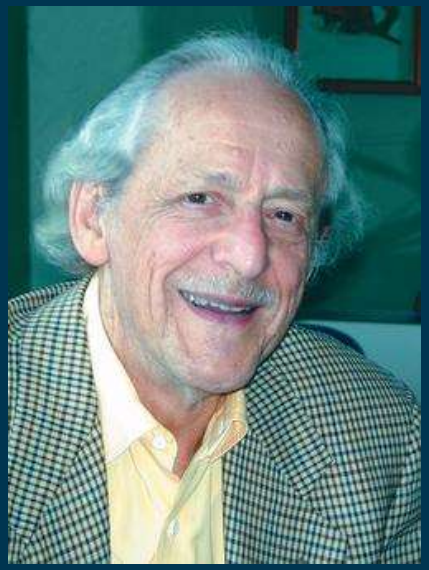
What kind of data would, in your opinion, be especially interesting/useful to enrich your database* with?

- Feel free to be creative! No wrong answers ☺
- Take a minute to think and write down some ideas
- Then discuss with your neighbours:
  - Do you have similar ideas? If not, would you add your neighbour's ideas to your list?
  - Do you think these ideas are practically feasible? Why (not)?

* 'your database' = either the one you work on/with in your daily practice, or the one your group has chosen as the focus

How?

Persistent identifiers are key



"For example, if each periodical would assign unique code numbers to the articles published, it would be possible for authors to list these numbers in their bibliographies and, thus, to save the work of coding on the part of the citation index staff. It is unlikely that such a development could take place in less than 5 or 10 years, but it is comparable to the problem of getting publishers to include Library of Congress card numbers in their publications."

– Eugene Garfield, 1955


isni
ORCID
doi
WIKIDATA
ISBN
GRID
ISSN

DOI

Identifier for (electronic version of) publication or other object

Controlled by a DOI registration agency, the oldest and largest of which is CrossRef

Who or what *is* CrossRef anyway?

- Started as an initiative of 12 journal publishers in 2000
- Set up as a non-profit organization, with membership of publishers and other DOI-using organizations
- Ech member can continuously update the metadata associated with their DOIs
- Currently ~11,000 members, ~109 million objects

Record

Identificatie	c:vabb:322868
Bookmark	https://anet.be/record/opacvabbg/c:vabb:322868/N

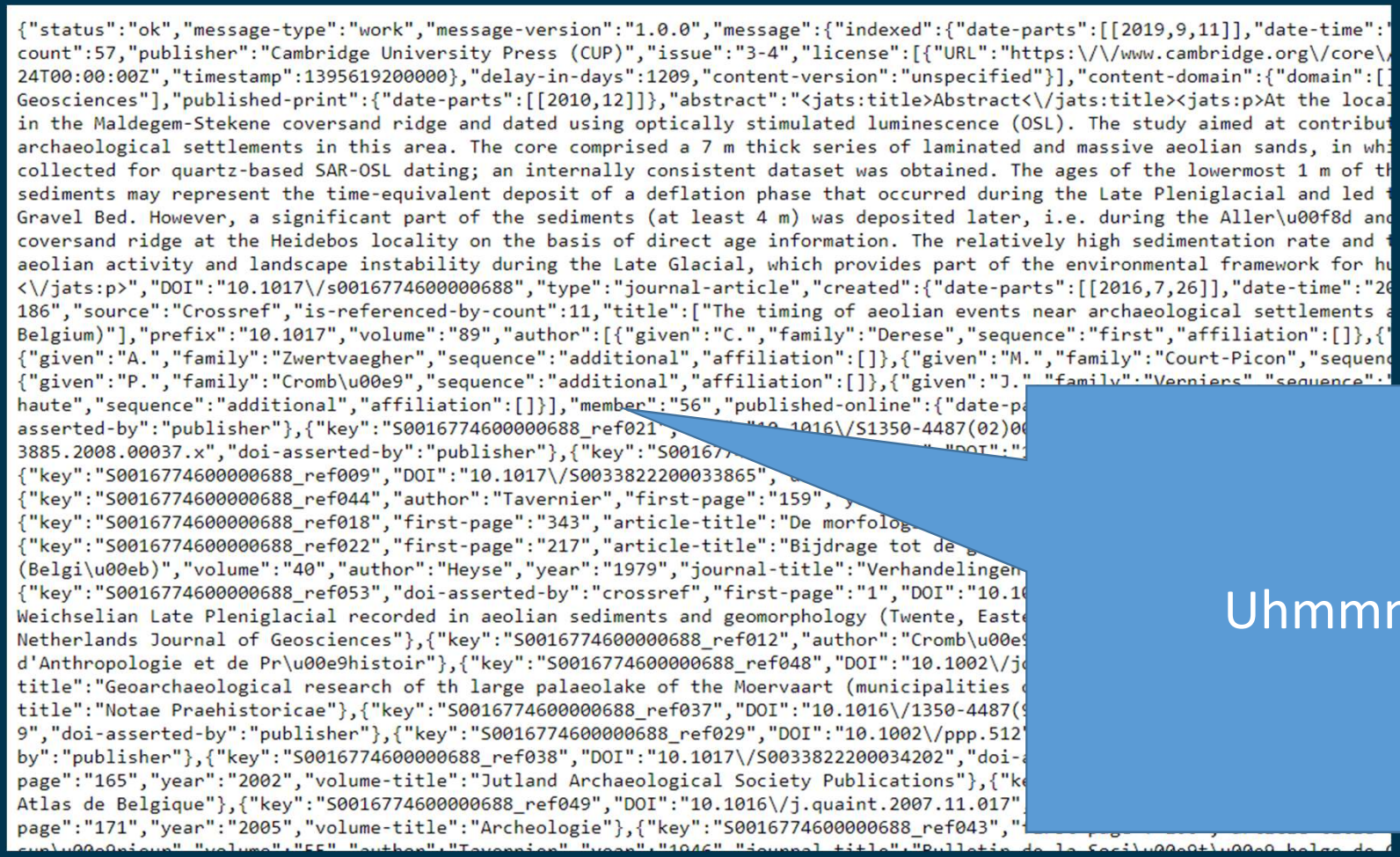
Publicatie

Titel	Economic hardship and depression across the life course : the impact of welfare state regimes
Auteur	Levecque, Katia Van Rossem, Ronan De Boyser, Katrien Van de Velde, Sarah Bracke, Piet
Nummer	Handle http://hdl.handle.net/1854/LU-1258270 doi 10.1177/0022146510394861 ISSN 0022-1465 ISI 000291611900008
Uitgave	2011
Omvang	52:2, p. 262-276
Annotatie(s)	Verkorte titel tijdschrift: J. Health Soc. Behav. published
In tijdschrift:	Journal of health and social behavior. - Los Angeles. - J. HEALTH SOC. BEHAV.. - . -
E-info	https://doi.org/10.1177/0022146510394861

Onderwerp

Collecties	VABB. Psychologie VABB. Sociologie VABB. Communicatiewetenschappen

https://api.crossref.org/works/10.1017/s0016774600000688



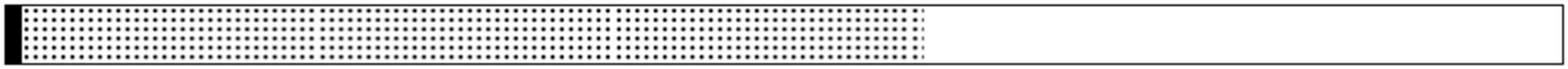
https://api.crossref.org/works/10.1017/s0016774600000688

https://api.crossref.org/works/10.1017/s0016774600000688


```
▼ reference:
  ▼ 0:
      key:                "bibr1-0022146510394861"
      DOI:                "10.1177/0952872002012002114"
      doi-asserted-by:    "publisher"
  ▼ 1:
```


How many citations are open today?

Initiative for Open Citations (I4OC)



1% 58% 41%


```
      volume-title:       "Poverty in Europe"
  ▼ 3:
      key:                "bibr4-0022146510394861"
      DOI:                "10.1177/1350506809341512"
      doi-asserted-by:    "publisher"
```

https://api.crossref.org/works/10.1017/s0016774600000688


```
link:
  0:
    URL:                  "http://journals.sagepub.com/doi/pdf/10.1177/0022146510394861"
    content-type:         "application/pdf"
    content-version:      "vor"
    intended-application: "text-mining"
  1:
    URL:                  "http://journals.sagepub.com/doi/full-xml/10.1177/0022146510394861"
    content-type:         "application/xml"
    content-version:      "vor"
    intended-application: "text-mining"
  2:
    URL:                  "http://journals.sagepub.com/doi/pdf/10.1177/0022146510394861"
    content-type:         "unspecified"
    content-version:      "vor"
    intended-application: "similarity-checking"
```

https://api.unpaywall.org/v2/10.1017/s0016774600000688?email=your@email.com


```
best_oa_location:       null
data_standard:          2
doi:                    "10.1177/0022146510394861"
doi_url:                "https://doi.org/10.1177/0022146510394861"
genre:                  "journal-article"
has_repository_copy:    false
is_oa:                  false
journal_is_in_doaj:     false
journal_is_oa:          false
journal_issn_l:         "0022-1465"
journal_issns:          "0022-1465,2150-6000"
journal_name:           "Journal of Health and Social Behavior"
oa_locations:           []
oa_status:              "closed"
```

https://api.altmetric.com/v1/doi/10.1017/s0016774600000688


```
altmetric_id:              582088
schema:                    "1.5.4"
is_oa:                     false
cited_by_posts_count:      3
cited_by_tweeters_count:   3
cited_by_accounts_count:   3
last_updated:              1421681839
score:                     1.75
history:
    1y:                    0
    6m:                    0
    3m:                    0
    1m:                    0
    1w:                    0
```

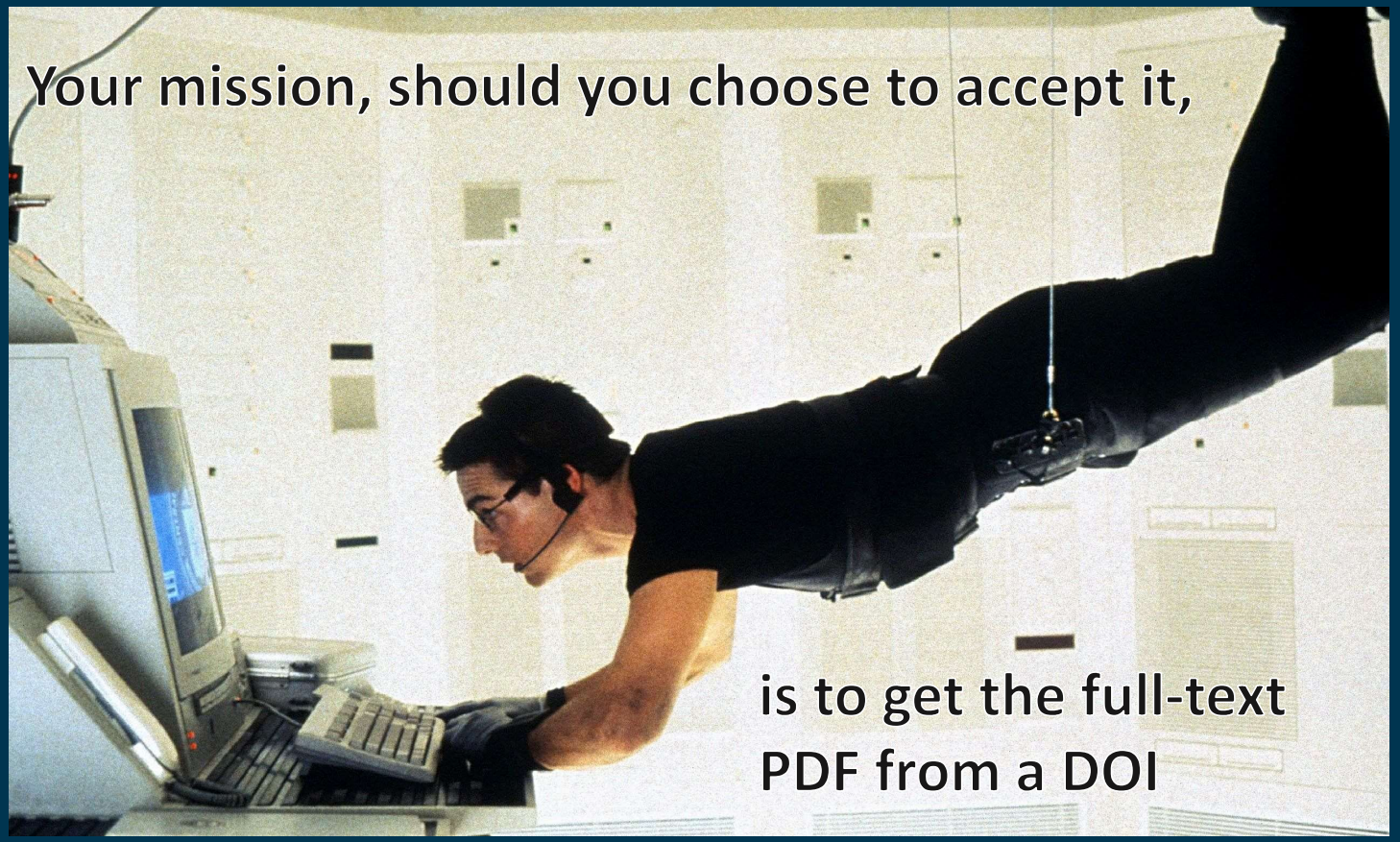

EASY
MEDIUM
HARD

Publicatie

Titel	The timing of aeolian events near archaeological settlements around Heidebos (Moervaart area, N Belgium)
Auteur	Derese, Cilia Vandenberghe, Dimitri Zwertvaegher, Ann Court-Picon, Mona Verniers, Jacques Van den haute, Peter
Nummer	Handle http://hdl.handle.net/1854/LU-1268703 ISI 000293365300001 ISSN 0016-7746
Uitgave	2010
Omvang	89:3-4, p. 173-186
Annotatie(s)	Verkorte titel tijdschrift: Neth. J. Geosci. published
In tijdschrift:	Geologie & mijnbouw. - Cambridge, [UK]. -
Korte inhoud	At the locality of Heidebos (Moervaart area, N Flanders, Belgium), a sedimentary core was taken in the Malde stimulated luminescence (OSL). The study aimed at contributing to an improved understanding of the evoluti in this area. The core comprised a 7 m thick series of laminated and massive aeolian sands, in which several samples were collected for quartz-based SAR-OSL dating; an internally consistent dataset was obtained. The

No DOI! Options?

The messy practice



The messy practice

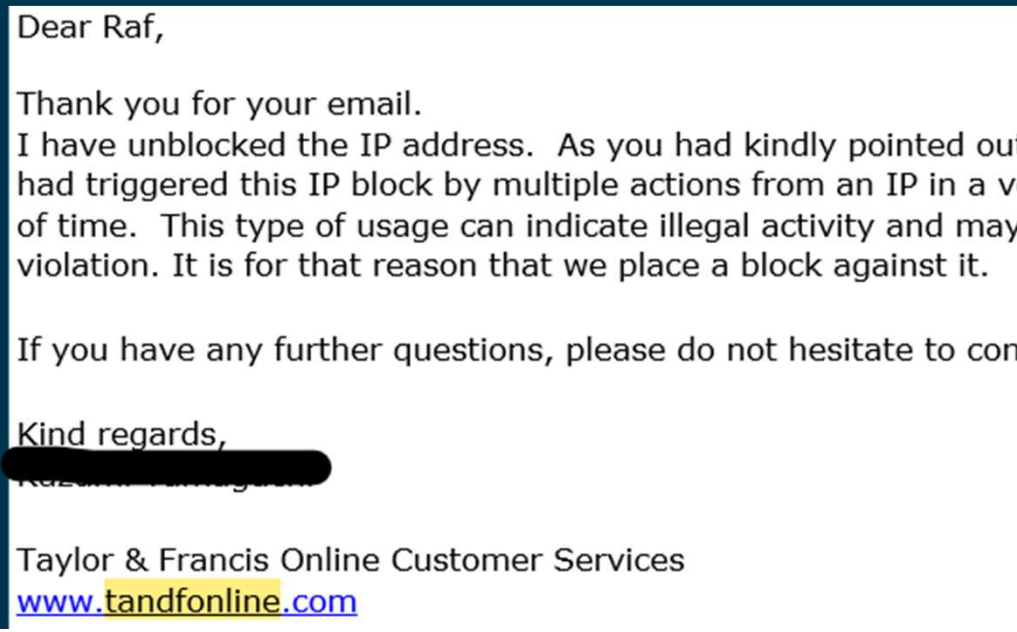
https://doi.org/*yourdoi* redirects to another document. This may be…

- the full-text PDF 🎉
- a web page that may, somewhere, contain a link to the PDF
  - <meta name="citation_pdf_url" value="*link_to_PDF*"> (Google Scholar)
  - Or try all links and see if anything PDF-like comes up…
- CrossRef metadata may contain URLs for text-mining (may require extra steps)
- Some publishers have a fairly straightforward URL scheme (e.g., https://www.jstor.org/stable/pdf/*yourdoi*.pdf)
- Or you can try, e.g., Unpaywall to find an open access PDF

The messy practice

You may also get:

- Other file formats than PDF: XML, e-books, images, plain text, and simply unreadable files…
- SSL errors
- empty pages
- timeouts
- unreachable servers
- blocked



Dear Raf,

Thank you for your email.
I have unblocked the IP address. As you had kindly pointed ou
had triggered this IP block by multiple actions from an IP in a v
of time. This type of usage can indicate illegal activity and may
violation. It is for that reason that we place a block against it.

If you have any further questions, please do not hesitate to con

Kind regards,
[redacted]

Taylor & Francis Online Customer Services
www.tandfonline.com

Enriching may introduce new problems

Number of publications with structured affiliation data by country of origin of the (co-)authors:

754

NUMBER OF PUBLICATIONS PER COUNTRY OF INSTITUTION OF AUTHORS

Country (48)	Number
Belgium	750
Netherlands	77
United States	22
United Kingdom	19

Enriching may introduce new problems

Alissa De Ceuninck is Master in Film Studies and Visual Culture at the University of Antwerp. Earlier, she obtained a Master in Communication Studies at the Free University of Brussels, specializing in media, democracy and journalism.

M. Morrens is connected as doctor training to be a psychiatrist to the Collaborative Antwerp Psychiatric Research Institute (CAPRI), Campus Drie Eiken, Antwerp.

Are these affiliations?
Why (not)?
Which institute(s) are they affiliated to?

More pitfalls

- External databases may be wrong while your database is right
  - Don’t blindly overwrite all values that are different!
- Not everything has a persistent identifier
  - Especially publications in local journals, publications in books or proceedings, gray literature…
  - VIRTA (Finland): 67% of peer-reviewed outputs has DOI, only 22% for book publications
  - Cuban journals cannot register DOIs because of trade embargo

We have new data. Now what?

Store all data in own database or fetch latest data from, e.g., CrosssRef when needed?

- Many data are dynamic, prone to change
- But fetching data when needed is only useful for ‘one at a time’ usage (e.g. showing some extra data in a record)

➔ Often storage with regular updates will be needed

We have new data. Now what?

"For databases where data are collected by means of data transfer, it is important to consider the relationship between the data in the national database and in the database from which the data originated. If the data are enriched in the national databases, it is useful to implement procedures that allow to improve also the accuracy of data in the databases from which the data originated. This, however, requires coordination between different organisations, consideration of the ownership of data as well as different legal frameworks that might influence this process."

- ➔ Implement feedback loop from national database to source database(s)
  - ➔ Not always possible for legal, technical… reasons
  - ➔ Extra validation by source organisations may be needed

Take-away messages

- Validating and enriching is useful but not always easy
- Persistent identifiers are core infrastructure to enable this. Objects that lack them are much harder to compare to external systems.
- Don’t blindly trust the external system for corrections.
- Be aware of dynamicity of data, i.e., think carefully about storage and ‘refreshing’.
- Implement feedback loop to source databases, if needed and possible.