The road towards structured affiliation information in a national bibliographic database

Peter Aspeslagh¹, Tim Engels, Raf Guns

Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp,
Middelheimlaan 1, 2020 Antwerp, Belgium

Abstract. The implementation of a Flemish research evaluation parameter highlights the complexity of author affiliation data collection for publications not included in major bibliographic databases. In this paper, we discuss a set of fundamental challenges that were encountered during a first data collection project. More specifically, we will elaborate the multifaceted data retrieval approach, the quest for a sustainable way of data registration and the development of necessary infrastructure and procedures. Although a lot of efforts are being invested in optimizing the exchange of well-structured author affiliation data, we will zoom in on opportunities that might arise to facilitate similar projects in the future.

Author affiliation data is an essential part of publication metadata. The more it becomes structurally available, the more it will be used in processes like research evaluation. However, the deployment of an affiliation data collection process for a Flemish research assessment parameter in social sciences and humanities demonstrates that continued efforts to optimize the quality and increase the distribution of affiliation data are necessary. Taking a diversity of publication practices into account, the creation of this Flemish dataset requires a multifaceted approach that not only includes the common data sources, but also an integrated set of ad hoc strategies beyond these tools. The organization of this operation, the challenges and opportunities will be elaborated in this paper.

1 Background

In the region of Flanders, Belgium, the University Research Fund (*Bijzonder Onderzoeksfonds – BOF*) is an inter-university performance-based research funding mechanism that distributes financial resources over the five different Flemish universities. Multiple parameters are applied, including a bibliometric indicator using the *Flemish Academic Bibliographic Database for the Social Sciences and Humanities* (VABB-SHW) – a compilation of research output of social sciences and humanities units at the five Flemish universities as main data source for calculations [1].

¹ Corresponding author: Peter.Aspeslagh@uantwerpen.be

With the reform of the BOF-regulation in 2019, a new parameter has been introduced. This parameter measures international collaboration through co-authorship. If a publication has two or more authors and at least one of them has a non-Belgian affiliation, the publication is considered the result of international collaboration and is counted towards this parameter.

The affiliation data as mentioned on the actual publication or on the (official) website of the publisher acts as primary data source. Contrary to databases like Web of Science (WoS) or Scopus, VABB-SHW does not (yet) include affiliation data.

However, almost half (48,1%) of all VABB-publications are also indexed in the Web of Science and possess a WoS Accession Number (UT-search field tag). As virtually all publications in the Web of Science contain affiliation data, matching via UT-tag quickly enriches the targeted dataset.

The other half of VABB-SHW is not included in the Web of Science. This means that for a 10-year timeframe, author affiliation data for all publications with at least one co-author needs to be collected retroactively in an alternative way by ECOOM (the Centre for R&D Monitoring), leaving us with a set of 19.440 individual publications. The data collection process for this subset will be discussed in this paper.

Because of the size of the dataset, we have decided to collect affiliation data at the level of publications, rather than publication-author-level. While the latter would enable some fine-grained analyses, it is not needed for the calculation of the BOF-key.

This process faces challenges on three fronts: (1) the (logistic) deployment of procedures, infrastructure and staff, (2) approaching sources and selecting methods for data retrieval and (3) defining a format for structural registration of affiliation data.

2 Operational context

Each of these challenges is framed within a broader context. The author affiliation data collection process is not only part of a funding procedure in a legal framework, but also serves bibliometric research and optimizes data quality of a national bibliographic database.

First, apart from the initial retroactive data collection operation, the process is recurrent: it needs to be repeated annually as each year a batch of a few thousand new publications is added to VABB-SHW. Due to administrative procedures a fixed timeline was proposed, which results in an annual data collection period of around six months.

Second, as the affiliation data is used as a data source in a performance-based research funding mechanism, validation by different actors is required. After the (annual) data collection period, random or full checks of the gathered author affiliation data are processed by the five Flemish universities. This is the main reason for the reduction of the data collection period: the funded actors need to be able to validate the affiliation data of their publications and propose alternatives if accuracy of selected affiliation data is disputed. Although the criteria for counting as international publication are well-defined, discussions might arise as the level of detail of the available author affiliation data varies. It only stresses the importance of unambiguous affiliation data.

Third, procedures to collect author affiliation data are not to be separated from the continuous data enrichment efforts of ECOOM. Although VABB-SHW contains sufficient metadata to meet its targets, bibliographic databases are never complete or fully accurate. Moreover, some data is already integrated into VABB-SHW before final bibliographic data is available. For example, journal articles published online before being included in a final publication count as valid publications for the BOF-parameter. Subsequently, some volume and issue numbers are missing in VABB-SHW. This affiliation data collection procedure, during which a subset of individual publications are consulted in one way or another, provide

a unique opportunity to fill these gaps. Infrastructure to register missing and update inaccurate bibliographic metadata needs to be available.

A successful implementation of the project requires not only efficient data retrieval, data registration and thorough infrastructure. It also has to respect deadlines, interact the data with other entities and optimize a national bibliographic database.

3 Infrastructure, procedures and staff

The project can only be successful if sufficient infrastructure and competent staff is available, guided by rigid procedures.

Integrating affiliation data from a multitude of sources, consistent registration, an up-to-date inventory of the subset of publications and an instance that allows validation require a central instrument that stores and coordinates the data. Therefore, an online platform and database was developed (Fig. 1). It consists of a copy of the bibliographic database (VABB-SHW), extended with functionality to enrich all components (publications, authors, journals, publishers) with diverse kinds of additional data, in bulk or manually via web forms.

In the application, a wizard-like coding form is used to register the author affiliations. It starts with an overview of all relevant bibliographic VABB-SHW data that is necessary to retrieve the document manually if affiliation data was unavailable in external bibliographic databases (cf. *infra*). A second tab contains web forms to store the affiliation data: after pasting full affiliation data mentioned in the byline or on the official web page of the publication, countries and institutions are normalized by manually assigning them via a dropdown menu and an autocomplete field. During a third step, accuracy of the DOI is checked or added if unavailable, followed by a form that registers if either a pdf file of the full text or a pdf print of the official web page is stored. An abstract can be checked or added in the fifth tab. The final component reviews the coding and enables staff members to confirm and validate the data about the particular publication. If no (sufficient) data is found, a 'check hardcopy'-button moves the record to a list for publications to be checked in a library.

The coding form of each publication is accessible via 'to do' lists that sort publications by journal, publisher or library where a hardcopy is available. Once a publication is coded, it disappears from the list.

Even if part of the affiliation data can be generated from external databases, all publications need a manual check (and validation). Specific procedures are stipulated, guidelines are distributed among staff members, who are intensively trained and supported by senior staff.

All of this turns out to be a labor-intensive process. Even if procedures can be generalized, the formulation of author affiliation data – if formulated – varies enormously as does the level of detail. Non-automated normalization requires interpretation, which, subsequently, requires expertise and extensive coaching, especially if new staff members arrive.

The platform is, apart from the affiliation data coding, fully embedded in the operational context of the organization. It contributes, on the one hand, to the data enrichment aspect by allowing to store additional or to correct current bibliographic metadata of VABB-SHW. On the other hand, it enables the five Flemish universities to perform author affiliation data checks and suggest corrections or additions in a separate application, connected to the central database.

1. Affiliation data (manually added or external sources) [c.vabb:469653]

You can register the affiliation data in this form. First, paste the affiliation data as available on the fulltext or on the website of the publication on the left side of your screen (A). Indicate if the points below apply. Subsequently, code the affiliation data by distilling the countries and affiliated organizations from the text you pasted (B). Finally, validate the data by pushing the 'Validate' button on the top right of your screen if authorized by the ECOOM staff.

A. Paste affiliation data & notifications

Save affiliation data

Paste affiliation data (textual)

Amandine Thiry is a PhD student in modern history at the Université Catholique de Louvain and Ghent University. Within the collaborative digital humanities project 'Tic Belgium', she investigates the transnational networks of prison experts in the period 1830–1914, with a focus on the role played by Belgian prison reformers.

Thomas D'Haenincx obtained his PhD in modern history at Ghent University in 2018. His doctoral research at Ghent University/Maastricht University is part of the collaborative digital humanities project 'TIC Belgium'. He studies the 'culture of international congresses' via a relational and actor-centred approach by applying SNA techniques to a large sample of congress attendees. He focuses mainly on congresses related to moral reform.

Christophe Verbruggen is an associate professor at the research unit 'Social History since 1750' at Ghent University and director of the Ghent Centre for Digital humanities. He is a coordinator of the project 'TIC Belgium', which focuses on the transnational dynamics of social reform in the Low Countries between 1840 and 1940. He specialises in the social history of intellectuals and cultural mobility in the 19th and 20th centuries.

Source

No affiliation data ☐

No affiliation data due to missing page in fulltext ☐

Partial affiliation data ☐

Guest lecturer among affiliations ☐

URL Publication

B. Code affiliation data

Assign country -- Belgium -- Add

Assign institution Add organization... Add GRID

Countries

Belgium Delete

Netherlands Delete

Organizations

Organization	Department	
Université Catholique de Louvain		Delete Update
Ghent University	Ghent Centre for Digital humanities	Delete Update
Maastricht University		Delete Update

Fig. 1. Author affiliation data collection coding form

4 Data retrieval: a multifaceted approach

The first step in the procedure described above, retrieving author affiliation data of VABB-SHW publications not included in the Web of Science, is carried out via two main channels: by querying external bibliographic databases if sufficient VABB-SHW metadata is available or by manual consultation of each of the publications. The former method returns quick gains trusting the accuracy of external databases, while the latter is time-consuming but allows the benefits of consulting primary sources. The combination of the data from external bibliographic databases and manual intervention constitute the final dataset.

4.1 Querying external bibliographic databases

External bibliographic databases are approached in two ways: either via a unique identifier (DOI or Web of Science UT code) or by title matching.

A unique identifier provides exact matches with records in bibliographic databases. These were used to query CrossRef, Scopus, MS Academic by DOI, and Book Citation Index (BkCI) and Emerging Sources Citation Index (ESCI) by UT code. However, two limitations have to be noticed: less than half of the targeted VABB-SHW records contain a DOI and structured affiliation data is not as omnipresent in some of these compared to the Web of Science. Even for recent publications in CrossRef, affiliation data is available for less than 25% of publications [2]. Still, matching via DOI or other identifiers remains a valuable effort.

Title matching was used to check if corresponding publications were found in Crossref, Scopus or MS Academic. As often multiple results were returned, a system with scoring criteria on several bibliographic variables was established.

Due to the complexity of the data collection operation, the querying of external bibliographic databases only took place at a later stage. In total, 10.909 publications were

matched with the external bibliographic databases, resulting in affiliation data for 2.218 individual publications. The data was normalized via automated scripts and integrated in the database.

4.2 Manual retrieval

The affiliation data for the majority of the publications needs to be collected manually. Therefore, two data retrieval tracks are activated: (1) registering the affiliation data from the full text and/or website of the publisher and (2) consulting hardcopies in an academic library one by one.

To facilitate the data collection, several strategies were developed to point staff members to the accurate documents to be used for coding, as reality turned out to be more complicated than originally estimated. Coaching is essential in order to identify the correct document as, apart from external bibliographic databases, only primary sources are allowed to be used for coding. Often affiliation data can be found in, for example, university library catalogues, but this might contain modified data as sometimes original author affiliations were converted to standardized internal organizational units. Staff members are also instructed to double check if the retrieved document was actually the one that they were looking for. Sometimes documents with the same title, authors and publication dates were published in other journals or as reports on faculty websites. More complicated cases were discovered in due course.

In order to collect as many correct data as possible, a priority order of data sources and document types was elaborated. The full text published on the official website of the journal/publisher is of first and foremost importance. If this is unavailable, Google Scholar is used to retrieve the full texts in online repositories or on academic social media profiles (ResearchGate, Academia, etc.). If, after a reasonable number of attempts no full text was found, affiliation data on the website of the journal/publisher can be used for coding. Additionally, data found in lists of contributors are considered to be valid as well. If, finally, no data is available online, the record is transferred to a list of publications of which the hardcopy needs to be retrieved.

It is self-evident that, in order to prove the provenance of the data, a pdf of the full text or the affiliation data on the website of the publisher needs to be saved. In the case of hardcopies, scans are required. A ScanTent was acquired using the DocScan app on a smartphone.

Efficiency gains were made by using data about publication access. As data collection takes place within the University of Antwerp network, information about subscription access (apart from Open Access publications) supported the creation of a priority list for publications to be coded.

5 Data registration

The registration of affiliation data needs to be as consistent as possible, using persistent identifiers (PIDs). Institutional identifiers (e.g. GRID, ROR) are only recently emerging in bibliographic databases like CrossRef. Even if the internationalization parameter is only a binary option (yes or no), extending VABB-SHW with affiliation data has to be done in an unambiguous (and reusable) way. Using fixed institutional identifiers like GRID or ROR greatly improves the value of the affiliation data and guarantees the sustainability. We opted to use GRID (*Global Research Identifier Database*) at the time of planning, as ROR was not yet available. GRID was developed by Dimensions as an entity list for organizations involved in research [3] and is composed of a set of almost 100.000 unique organizations.

The data about the affiliated institutions that is retrieved, both by querying external databases or via manual intervention, is converted to a GRID identifier.

However, the fluid nature of research organizations and educational institutions demonstrates a weakness: it doesn't always allow a straightforward connection between the institution on the publication and a corresponding one in the database of institutions. Names are constantly changing, organizations are merging or disappearing. A temporal dimension in an international database is not always up to date.

Moreover, local contexts are less extensively captured in (Anglo-Saxon oriented) international standardization operations. For example, some countries can have multiple governmental layers, which complicates standardization if scientific institutions are linked to one or more layer(s). In a federal country like Belgium many such issues arise.

This had to be addressed during the development of the database and the application. In order to code the affiliated institutions, the full GRID dataset was imported and an internal identifier was assigned. But in order to code *all* the institutions mentioned on the full text or the official web page of the publications, we needed to have the opportunity to add additional institutions, beyond the GRID dataset. This was realized and resulted in a database of GRID- and non-GRID-organizations containing the same variables.

After almost three quarter of the 19.440 publications were processed, we identified a total of 3.411 different institutions to which at least one author is affiliated, originating from 119 countries. Of those institutions, 2.277 could be linked to a GRID identifier (66,75%); 1.134 extra institutions were added to our database. This means that one third of the institutions to which authors were affiliated were not included in GRID (Fig. 2). Almost half of those additional institutions were from Belgian origin. In absolute numbers, only a quarter of the Belgian institutions were captured by GRID, contrary to, for example, institutions from the US or the UK.

Fig. 2. Number of GRID versus non-GRID institutions by country

We plan further research to get a better view on the institutions that are not captured by an international database, taking into account that the publications we coded are largely beyond the current scope of international bibliographic databases.

6 Opportunities

These three major challenges associated with affiliation data collection point to opportunities for future developments, not only to facilitate data collection operations, but also to propagate research and expertise.

6.1 Uniform and machine-readable author affiliation metadata scheme

We encourage the (continued) development of a uniform and machine-readable metadata scheme for affiliation data, published on the official web page of the publication and registered in systems like CrossRef. The inclusion of organizational identifiers (GRID, ROR) is a *conditio sine qua non*. In recent years, the Initiative for Open Citations (IIOC) and Initiative for Open Abstracts (IOA) have successfully advocated in favour of openly available references and abstracts. In a similar fashion, we recommend publishers to make affiliations openly available as FAIR data in, e.g., CrossRef.

6.2 Universal open publication of affiliation data

Even if a uniform and machine-readable metadata scheme is desirable, the mere open availability of author affiliation data would already considerably smoothen this data collection operation. In many cases affiliations are not printed on the full text or displayed on the website of the publication, nor does an (edited) volume release a list of contributors. We are aware that each journal or publisher may apply a well-thought policy for the content that is displayed on its website or on full texts, but the substantial number of publications that do contain accessible data prove that a systematic description of author affiliations is feasible.

Consulting services like Google Books sometimes solve the issue, but often pages containing the actual affiliation data are missing or are only partly available. If all pages with affiliation data would be included in the preview of a publication, another gap would be filled.

6.3 Simulating assigning DOIs

We recommend stimulating publishers or journals to assign a DOI to (and register relevant metadata for) each of their publications, even if they only have a national scope. Well-targeted communication highlighting the importance of the unambiguous publication of author affiliation data, based on shared metadata schemes and international nomenclatures, would considerably speed up a data collection process. Moreover, redirection via DOI increases visibility.

6.4 Local context

Large international organizational databases would benefit from national antennas, keeping the global identification list in correspondence with often complex national realities, similar to, for example, national experts of ERIH PLUS [4]. The dynamic nature of organizational entities requires constant supervision.

Discussion

It was our goal to illustrate the complexity of an author affiliation data collection project for publications that are not easily captured by common bibliographic databases. We focused on the practical challenges it entailed and on opportunities that could facilitate future projects.

There's no quick solution to all the issues that we summed up. Some of them were already addressed in the course of the project: additional data sources emerged, some procedures were shortened or recalibrated, the online application was updated whenever improvements could be made etc. This requires a substantial degree of flexibility of everyone involved to keep the data collection as efficient as possible. We experienced, for example, that applications built in-house can quickly adapt to changing circumstances, preventing long delays. However, a broader reflection on good practices for (author affiliation) data collection can improve processes.

It is also clear that this is more than just about developing identifiers, standards and systems; it's also a matter of raising awareness. For example, increase the availability of author affiliation data on websites of publishers or stimulating the assignment of DOI's does not automatically require the creation of new tools, but rather the continued implementation of existing instruments. Diverse initiatives to promote accessible affiliation data are contributing to a more comprehensive knowledge of publication practice and are beneficial for both research and civil service. It could be an incentive to use author affiliation data for the development even more precise bibliometric research evaluation indicators.

References

1. Engels, T. E., & Guns, R. (2018). The Flemish Performance-based Research Funding System: A Unique Variant of the Norwegian Model, *Journal of Data and Information Science*, **3**(4), 45-60.
2. Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata, *Quantitative Science Studies*, **1**(1), 414-427.
3. Hook, D.W., Porter, S.J. & Herzog, C. (2018). Dimensions: Building Context for Search and Evaluation. *Frontiers in Research Metrics and Analytics*, 3:23.
4. Vikanes Lavik, G.A., & Sivertsen, G. (2017). ERIH PLUS – Making the SSH Visible, Searchable and Available. *Procedia Computer Science*, **106**, 61-65.