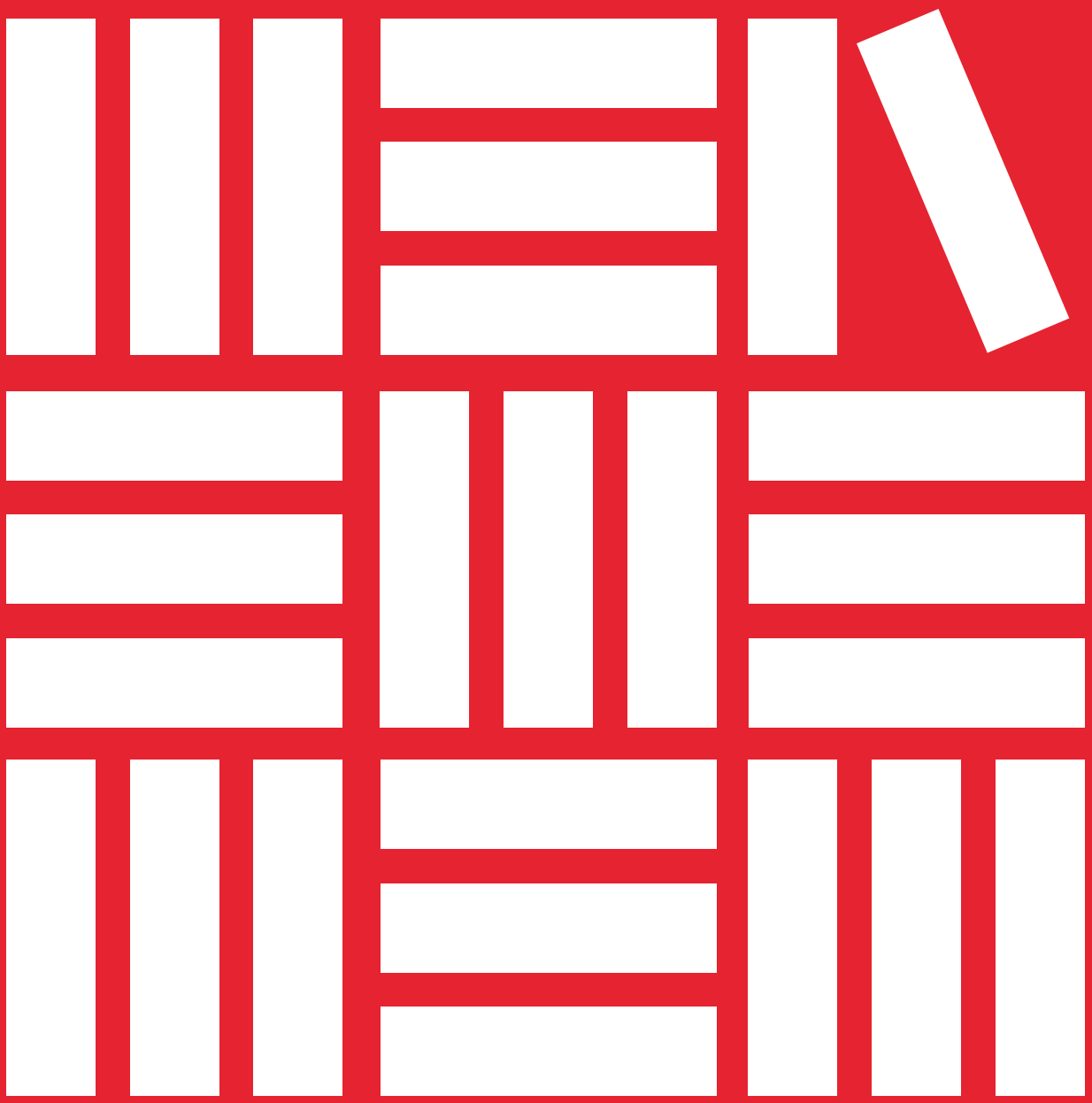

CREATING AND MAINTAINING A NATIONAL BIBLIOGRAPHIC DATABASE FOR RESEARCH OUTPUT

Manual of good practices

ENRESSH & ECOOM

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
ENRESSH
ecoom

CREATING AND MAINTAINING A NATIONAL BIBLIOGRAPHIC DATABASE FOR RESEARCH OUTPUT

Manual of good practices

Linda Sīle

Centre for R&D Monitoring (ECOOM), University of Antwerp, Belgium

Raf Guns

Centre for R&D Monitoring (ECOOM), University of Antwerp, Belgium

Dragan Ivanović

University of Novi Sad, Serbia

Janne Pölönen

Federation of Finnish Learned Societies, Finland

Tim C.E. Engels

Centre for R&D Monitoring (ECOOM), University of Antwerp, Belgium

ENRESSH & ECOOM: 2019

About ENRESSH

The “European Network for Research Evaluation in the Social Sciences and the Humanities” (ENRESSH, www.enressh.eu) is a COST Action, starting in April 2016 and ending in April 2020. ENRESSH aims to propose clear best practices in the field of SSH research evaluation. The Action brings together more than 125 experts from 36 countries, such as researchers in evaluation studies, policy makers and members of evaluation units, as well as researchers from SSH disciplines. Its approach is based on the comparison and the cross fertilisation of strands of work dedicated to SSH research evaluation, currently under development in different parts of Europe, seeking to avoid unnecessary duplication and to upscale results.

Acknowledgements

The authors thank all members of ENRESSH who engaged in fruitful discussions about and provided valuable feedback on earlier versions of this manual. Similarly, we are indebted to the numerous participants in the two ENRESSH surveys on national databases for research output in the social sciences and humanities and the workshop ‘Working with national bibliographic databases for research output’. Awareness of the wide range of database designs that resulted from those interactions has been the point of departure for this manual. Special thanks to Hanna-Mari Puuska and Jadranka Stojanovski for their helpful suggestions.

LS, RG, and TE thank the Flemish Government for its support to the Centre for R& D Monitoring (ECOOM).

Suggested citation: Sīle, L., Guns, R., Ivanović, D., Pölönen J., and Engels T.C.E. (2019). Creating and maintaining a national bibliographic database for research output: manual of good practices. ENRESSH & ECOOM: Antwerp. DOI: 10.6084/m9.figshare.9989204

CONTENTS

INTRODUCTION

Scope and audience

In this manual of good practices for the maintenance of national bibliographic databases for research output, we outline several aspects that we regard as especially important for databases that are used for evaluation and funding allocation purposes. Databases can be implemented in numerous ways. The choice of implementation depends on the envisioned uses as well as on the available resources and expertise. A bibliographic database can be a stand-alone system created for the sole purpose of funding allocation, or it can be a module linked up with other modules in a Current Research Information System addressing various purposes. The technical solution can be developed in-house or it can be an open-source or commercial solution. In other words, each database is unique. Hence it is neither feasible nor desirable to propose one set of good practices that everyone ought to follow. What works well in one context, may not work in the same way when transferred to a different context. Therefore we highlight 30 issues that are worth considering in database design, organisation, maintenance, and usage. However, we do not claim that all of them are of equal importance for all databases.

This manual has been developed in the context of the 'European Network for Research Evaluation in the Social Sciences and Humanities' (ENRESSH, COST Action 15137, www.enressh.eu). The manual is intended as an informative resource and a trigger for discussions for people who are involved with the implementation and maintenance of national bibliographic databases for research output. This includes developers, database administrators, policy makers, librarians, researchers, and everyone else who is in one way or another involved in database work. This manual builds on earlier work carried out in the frame of ENRESSH and related to national bibliographic databases for research outputs, in particular the overview and analysis of such databases (Sīle et al 2017, 2018), and the VIRTA-ENRESSH pilot of a European database for research output in the social sciences and humanities (SSH) (Puuska et al 2018).

Although all of the issues we highlight are applicable to generic databases or bibliographic databases for any knowledge domain, our focus has been on requirements for databases specific to SSH. Hence in some recommendations, one will find more emphasis on this knowledge domain. All suggestions are written anticipating the use of this manual at different stages in the implementation of national bibliographic databases for research output. Issues we describe are intended to be relevant for databases that are currently being designed or developed, as well as for databases that have been operating for many years.

Background

Over time, bibliographic databases for research output have become an important tool and source of insights in research evaluation and funding allocation activities (Jonkers and Zacharewicz 2016). However, the use of bibliographic databases for these ends comes with certain requirements.

In the recent decade, several initiatives, like the Leiden Manifesto (Hicks et al. 2015) and the San Francisco Declaration on Research Assessment (DORA 2012) have outlined principles to be followed in research evaluation activities. With respect to data, both of these initiatives foreground the need for openness and transparency in data collection processes: "construction of the data-

bases required for evaluation should follow clearly stated rules, set before the research has been completed" (Hicks et al. 2015); see also principle 11 in DORA). Moreover, it is proposed that data collection should involve researchers whose research will be evaluated: "To ensure data quality, all researchers included in bibliometric studies should be able to check that their outputs have been correctly identified. Everyone directing and managing evaluation processes should assure data accuracy, through self-verification or third-party audit" (Leiden Manifesto, principle 5). Finally, most importantly, both documents emphasise the need to acknowledge the diversity of research practices that can manifest in numerous forms of publications and research outputs (principles 3 and 6 in the Leiden Manifesto and principles 3, 5, and 14 in DORA). This latter point is especially important for SSH where researchers use a wide range of media to communicate their research (Nederhof 2006; Hicks 2004). The same applies to some medical fields, computer science, engineering, and other knowledge domains. National bibliographic databases are especially valuable as they tend to be of much wider scope than international bibliographic databases which typically do not have good coverage of SSH (Kulczycki et al. 2018). Typically these databases include not only metadata on journal articles, but also on books, chapters, textbooks, newspaper articles and other publications and artifacts. All these considerations guide our thinking about good practices for national bibliographic databases for research output.

Another source we have drawn on is the FAIR data principles (Wilkinson et al. 2016). Embedded in the Open Science movement, the FAIR principles were developed to improve research data management practices that are necessary for sharing and reuse of data. With the increasing use of bibliographic data for research evaluation and monitoring purposes, the FAIR data principles are applicable in this context as well.

Within the framework of FAIR data, data need to be Findable, Accessible, Interoperable, and Reusable. Each of these four principles is accompanied with more detailed specifications (e.g. for findability: 'F1. (meta)data are assigned a globally unique and persistent identifier', Wilkinson et al. 2016). These considerations strongly resonate with the call for openness and transparency in research evaluation activities one can find in DORA and the Leiden Manifesto. Hence the FAIR data principles are equally relevant for bibliographic metadata—data used in research evaluation and funding allocation settings.

Definitions

Through the development of this manual, we have encountered a wide variety in the terminology that is used to describe different aspects related to bibliographic databases for research output. Often the same aspect is described using different terms depending on the context of its use: librarians prefer one term, bibliometricians use another, and developers are more familiar with yet another. It is beyond the scope of this manual to try and define all terms. Yet we hope that the following definitions of the main terms will suffice to communicate the ideas we have gathered.

Bibliographic database: a structured set of bibliographic metadata records. We use this very broad definition since we acknowledge that technical solutions that are used to store records of bibliographic metadata can considerably vary in their complexity. Consequently, this manual is intended to be equally applicable to complex information systems and to basic databases. Similarly, the practices described here can also be of use for digital repositories, archives, and other similar systems.

Research output: scholarly publications and other artifacts (e.g., corpora, works of art, performances, software, exhibitions) that represent or communicate research findings. This broad defi-

nition, in some contexts, might also include research data. However, given that research data have additional requirements for metadata curation and most national bibliographic databases currently in operation do not include research data, we consider research data to be beyond the scope of this manual.

Data and metadata: for the definition of data (plural from datum) we adopt (with a slight adjustment) a definition proposed by Christine Borgman 'representations of observations, objects, or other entities used as evidence of phenomena' (Borgman 2016, 28). Metadata commonly are understood as 'data about data' or, in other words, data that are used to describe and characterise data and relationships among them (Borgman 2016). However, we need to highlight that in the area of bibliometrics, research evaluation and bibliographic databases for research output, the two terms sometimes get to be used interchangeably. As noted, the content of bibliographic databases is bibliographic metadata that refer to research output.

Structure

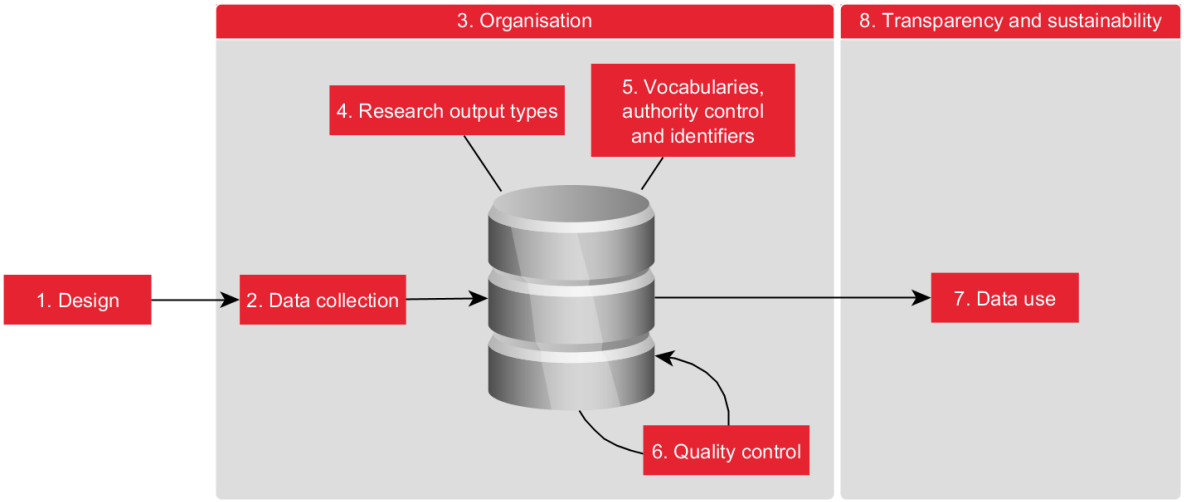

Figure 1 Structure of the manual

The structure of this overview is based around 8 themes, namely Design; Data collection; Organisation; Research output types; Vocabularies, authority control and identifiers; Quality control; Data use; Transparency and sustainability (Figure 1). Selection of these themes is based on experience and discussions of those involved in the maintenance of bibliographic databases for research output. In each theme, we make a number of statements highlighting practices which we regard as worth learning from. For each theme, we highlight the main benefit of implementing such a recommendation along with a brief description of possible ways of introducing the recommendation in practice. At the end of the manual, one can find contact information and links to the websites of a number of national bibliographic databases for research output.

References

Borgman, C. L. (2016). Big data, little data, no data: Scholarship in the networked world (Cambridge, Massachusetts; London, England). MIT Press.

DORA. (2012). San Francisco Declaration on Research Assessment. Retrieved from https://sfdora.org/read/

Hicks, D. (2004). The Four Literatures of Social Science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), Handbook of Quantitative Science and Technology Research (pp. 473–496). Dordrecht: Kluwer Academic Publishers.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. Nature, 520, 429–431.

Jonkers, K., & Zacharewicz, T. (2016). Research Performance Based Funding Systems: A Comparative Assessment. Retrieved from European Commission: Joint Research Centre website: https://rio.jrc.ec.europa.eu/en/file/9514/download?token=-8JG6aKx

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., … Zuccala, A. (2018). Publication patterns in the social sciences and humanities: Evidence from eight European countries. Scientometrics, 116(1), 463–486. https://doi.org/10.1007/s11192-018-2711-0

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review. Scientometrics, 66(1), 81–100. https://doi.org/10.1007/s11192-006-0007-2

OECD. (2015). Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development,. https://doi.org/10.1787/9789264239012-en

Puuska, H.-M., Guns, R., Pölönen, J., Sivertsen, G., Mañana-Rodríguez, J., & Engels, T. (2018). Proof of concept of a European database for social sciences and humanities publications: Description of the VIRTA-ENRESSH pilot (p. 23). Retrieved from CSC & ENRESSH website: https://doi.org/10.6084/M9.FIGSHARE.5993506

Sīle, L., Guns, R., Sivertsen, G., & Engels, T. C. E. (2017). European Databases and Repositories for Social Sciences and Humanities Research Output (p. 25). Retrieved from ECOOM & ENRESSH website: https://doi.org/10.6084/m9.figshare.5172322.v2

Sīle, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T. C. E., Arefiev, P., … Teitelbaum, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. Research Evaluation, 27(4), 310–322. https://doi.org/10.1093/reseval/rvy016

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

DESIGN

1. Identify and make explicit the purpose(s) of the database

Helps to design the database in line with users' needs

When designing a new database, purposes can be identified by means of a consultation and a discussion with all relevant users and stakeholders. Such discussions can be informed by examples of existing databases (in one's own country or abroad) and their usability with respect to different purposes. In relation to an existing database, the identification of purposes is a process of reflection and explication of principles that has guided existing database work. It can be that multiple purposes exist side by side. In such cases, it is useful to prioritise one or several purposes. Having a clear and explicit purpose or set of purposes, known to all stakeholders, makes it possible to guide the following steps and efficiently communicate about them.

2. Draw on expertise from all relevant knowledge domains

Enables tackling the complexity of design, setup, and organisation of a national database with broad, relevant and up-to-date expertise

For the specification of database design, it is crucial to take into account the views of researchers, librarians, bibliometricians, and policy makers. For the technical solution, one requires developers and database administrators. Similarly expertise from library practice as well as research administration is beneficial for the specification of metadata structure, expertise in cataloguing, research information systems. Furthermore, additional expertise from other domains might be required depending on the specifics of the envisioned database. If one of the knowledge domains is not taken into account when implementing or maintaining the database, there is a risk that problems that inevitably emerge in database work will not be solved using the most up to date relevant expertise.

3. Define the data model and/or metadata schema, taking into account the database's purpose and recognized standards

Ensures that the system can fulfil its purpose, while following recognized standards simplifies the work and can benefit interoperability

Prior to technical implementation, it is useful to make the database's data model explicit and obtain approval from all relevant stakeholders. The database's purpose should be the primary guiding

factor in making decisions, e.g., which entities are relevant, which metadata fields to include or exclude, or which fields should be mandatory. In addition, active collaboration and coordination of the work by experts from different knowledge domains is necessary to ensure that all features of the database design are aligned with the database purpose(s).

Where possible, it is advisable to make use of recognized standards for the data model, such as the Common European Research Information Format (CERIF) or an adaptation, like the European Publication Information Infrastructure Data Model[1] . Furthermore, there are several recognized standards relating to metadata and bibliographic records – both at the level of data contents, such as Resource Description and Access or the International Cataloguing Principles, and at the level of data structure, like Dublin Core or Metadata Object Description and Schema (MODS) – that can be adapted partially or completely. Note, however, that most bibliographic standards do not take evaluation (or funding allocation) purposes explicitly into account. If the database is intended to be used for such purposes, make sure that the data model contains all necessary fields (e.g., affiliation data for authors).

4. Select a suitable technical solution and design the technical structure of the database

Contributes to the functionality, performance, and maintainability of the database

The purpose of the database should be the central factor in deciding which technical solution is chosen. Additional factors include the available budget; the estimated number of records and requests; contemporary technologies/databases and their characteristics; and prior experience of staff (technicians and librarians) with certain technologies. It is useful to translate the purposes and needs in a detailed specification of requirements. However, one should avoid redefining the purpose to fit technological choices made for other reasons. These considerations apply both when implementing a ready-made CRIS system (e.g., DSpace-CRIS, Pure by Elsevier) and when building a new system.

Having chosen a technical solution, the data model needs to be translated into the technical structure of the system. Different (types of) database management systems offer different trade-offs and features, all of which may affect technical choices as well as design decisions.

[1] Nikkanen, Joonas (2019). Summary of European Publication Information Infrastructure Data Model. retrieved from https://wiki.eduuni.fi/display/cscvirtajtp/Summary+of+European+Publication+Information+Infrastructure+Data+Model

DATA COLLECTION

5. Collect the data systematically

Ensures that the database content is suitable for research evaluation and other uses where the comprehensiveness of data is one of the requirements

After the scope of the content has been delineated, design a workflow that will lead to the maximum coverage. For example, if the scope is limited to three output types from all state universities, agreements have to be made with each university on data transfer or other means of data collection at institutional level. Ideally, these agreements should specify deadlines when data will be provided and procedures that will ensure the completeness at individual level (e.g., mandate to report).

6. Avoid manual input where possible

Contributes to the efficiency in the maintenance of a national database

Where possible, use data transfer and avoid manual input. Manual input is a time-consuming process, while necessary data often already exist in some database or system (e.g., institutional repository or other national database). Even if the metadata format or the quality of data is not at the highest level or the data structure is not in line with the purpose of the national database, data structure can be adjusted and data quality can be improved. It might appear that data reuse is a cumbersome process both in terms of the technical challenges as well as the coordination that is necessary when working with data from multiple systems. However, the ability to achieve compatibility and/or interoperability between the national database and other systems is valuable expertise that can be applied when extending the scope of the national database or pursuing similar efforts with respect to other databases and information systems.

ORGANISATION

7. Collaborate with stakeholders

Contributes to usability, publicity and quality of the database

The identification of stakeholders, communication, and active collaboration are preconditions for a successful operation of a national bibliographic database for research output. Stakeholders here mean representatives of organisations involved in the implementation and the maintenance of a database, all groups of users of a database and researchers whose research output is represented in a database. To achieve this, a first step is to identify all persons, organisations and their representatives that are involved in the operation of the database as data providers, administrators, users or in any other role. It is crucial to treat also researchers whose research output is recorded in the database as stakeholders. If there is no collaboration with stakeholders, there is a risk of negative perceptions and unwillingness to participate in the operation of a national database.

8. Specify roles and responsibilities in the maintenance of the database

Increases the efficiency and contributes to the transparency of the database maintenance

When multiple different organisations are involved in database workflows, the coordination of different tasks can easily become complicated. For a smooth operation of the database, it is necessary to specify each step of the workflow along with the actor that is responsible. Particular attention has to be paid to steps that require frequent and repeated import/updates from other databases and to steps where disagreement about some quality-related aspect can occur. At these steps, it should be clarified in what way agreements should be made and who should be involved in this process.

9. Embed the national database in a national legal framework

Ensures the stability for the database

The need and the justification of a national database for research output should be embedded in a national legal framework (on scientific activity, research funding, etc.). This means that there is a commitment for the maintenance of the database at the government level. This, first of all, acts as an incentive for all involved actors. Secondly, this is a formal indication of the value of the different uses that a national database for research output enables. If the legal framework is too rigid, however, it may also hinder further development of the database – ideally, some flexibility is allowed.

RESEARCH OUTPUT TYPES

10. Aim for inclusion of a wide range of research output types

Facilitates multiple uses of the database

More detailed specification of the range of research output types depends on the purpose as well as the characteristics of research activities in a specific context. First of all, the bibliographic database should include all relevant types of research output. It should also allow distinction between output types, including articles in journals, articles/chapters in books, monographs and edited work, articles in conference proceedings. Ideally, the content should not be limited to scholarly publications, but also contain other publication and output types. Similarly, it is crucial to ensure that output from all relevant researchers and institutions are included in a database. However, the wider the scope, the more resources can be needed for coordination, data processing, and data input (in case data are entered manually). If resources for the maintenance of a database are limited, this can affect the quality of data and, consequently, the usability of the database.

11. Take into account characteristics of research output in different academic disciplines

Enhances the visibility of scholarship in all areas of research

Classification of output types that corresponds with practices characteristic to specific academic disciplines enables rich and detailed insights on the diversity in research practices. Some disciplines are focused more on communication in international peer-reviewed journals, while in others, it is important to publish in national journals, to prepare educational resources and textbooks, or to produce clinical guidelines. In some contexts, it can be sufficient to use a generic category 'scholarly monograph', but in others it can be more appropriate to make distinctions between different types of scholarly monographs (e.g. loose-leaf publications in law). The same applies to other research output type categories.

This can be achieved with a classification of research output types that is created in consultation with researchers in different academic disciplines. Ideally, all output types that researchers propose should be included either under a generic category or as a specific output type. Also, it has to be anticipated that the classification will require updates when new output types emerge.

VOCABULARIES, AUTHORITY CONTROL AND IDENTIFIERS

12. Maintain authority lists for publication channels

Contributes to the accuracy of data on publication channels and the functionality of the database

By 'publication channels' we mean journals, publishers, conferences, and other channels that support scholarly communication. Maintain dynamic authority lists of journals and publishers, based on cataloguing conventions. An authority list should contain a local persistent identifier and external identifiers (e.g., International Standard Serial Numbers (ISSN) for journals, International Standard Book Number (ISBN) prefixes for publishers), as well as other basic information such as the title and its variants. There are multiple ways to structure such lists. For journals, each unique ISSN can be represented by exactly one unique record or, in contrast, each unique journal (possibly corresponding to multiple ISSNs) is represented by exactly one unique record. Regardless of the choice, it is important to consistently use only one approach throughout all records in the authority list.

Depending on the database's purposes, additional information can be included in the authority record, such as information on peer review, perceived quality and impact, or open access status. Given the dynamic nature of both journals and publishers (mergers, splits, acquisitions etc.), it is useful to be able to specify relations between records and to have information on when a new authority record is started or the time frame for which the information is valid.

13. Maintain authority lists for authors and organisations

Contributes to the accuracy of data on authors and organisations and the functionality of the database

Maintain dynamic authority lists of authors and organisations, based on cataloguing conventions. Each author (organisation) should be represented in the database as a record with a persistent identifier. If other identifiers (institutional, national or international) are available, they should also be stored in this record. An author or organisation record should contain at least name (including variants) and, ideally, information on the time frame for which this information is valid.

For research evaluation and other uses it is important to structure the data on research output by author or organisation. This requires homogeneous identification of authors and organisations for each record of research output. Keep in mind that some systems make a distinction between authors (organisations) within the database's scope (e.g., from the country covered by the database) and those outside of the database's scope (e.g., from other countries). For the latter, authority control is more difficult and some systems treat them as uncontrolled text strings. If possible, authority control applies to all authors and organisations.

14. Use international persistent identifiers where possible

Increases interoperability with other national and international databases and systems

A number of information that is stored in a national database for research output can be represented by an international and globally unique persistent identifier. For digital publications and other research outputs, it is increasingly possible to use Digital Object Identifier (DOI). For journals and book publications one can use ISSN and ISBN respectively. For authors, it is possible to use Open Researcher and Contributor ID (ORCID), Virtual International Authority File (VIAF), International Standard Name Identifier (ISNI), and/or other international identifiers. For organisations, the work on international persistent identifiers is ongoing, but some options to consider are ISNI, Research Organisation Registry (ROR), and Global Research Identifier Database (GRID).

15. Use as much as possible terms from well-known and standardized vocabularies

Enhances the interoperability and functionality of the database

For languages' and countries' codes there are ISO 639 and ISO 3166 well-known and standardized vocabularies, respectively. Also, there are standardized vocabularies for scientific fields (e.g., OECD Fields of Research and Development, OECD FORD) and publication types (e.g., Consortia Advancing Standards in Research Administration Information, CASRAI). For specific purposes, it may be necessary to use local vocabularies and map part of them to standard international vocabularies, or even to define a completely new local vocabulary in the absence of a standardized vocabulary for a particular domain.

Adoption of standardized and well-known vocabularies will enhance adoption of a certain system by users. Also, in the case of interoperability with other systems, adoption of those standardized and well-known vocabularies can increase the amount of automated process, whether the source and target systems use the same vocabulary, or use different standardized vocabularies whose mapping of terms has been already defined.

16. When developing own vocabulary, consult stakeholders and relevant experts

Ensures that the vocabulary is usable and captures all use cases

Vocabularies can be entirely developed 'from scratch' or can be based on already existing vocabularies. For example, it is possible to use the CASRAI typology of research output types and complement with additional categories that are relevant for researchers and other stakeholders. Therefore, it is crucial that the development or adjustments of vocabularies should be carried out in consultation with all relevant stakeholders. Each value should be named, defined, and, in addition, it should be accompanied with guidelines for implementation. For example, guidelines could specify who determines and how the academic discipline of research output. The same considerations apply for research output types and other vocabularies.

QUALITY CONTROL

17. Implement a high quality deduplication procedure

Enhances data quality by avoiding the problem of multiple records for the same entity

Most national databases ingest data from multiple sources (e.g., automatic import from third-party systems, manual input by authors or other staff, data transfer from different institutions). The associated risk is that the same publication, author or organisation enters the database more than once as separate records. Such duplicate records make information retrieval more difficult and are the cause of inaccurate analysis and statistics. Deduplication refers to procedures that can track down and resolve possible duplicates. Resolution can happen in two ways: either only one record is retained (adding information from the deleted record where necessary) or both records are kept but linked.

The most straightforward way to identify multiple records of the same entity is to use persistent identifiers. If two records carry the same persistent identifier (e.g., DOI for publications or ORCID ID for researchers), then they refer to the same entity. Often, however, one cannot rely only on such identifiers. Advanced deduplication procedures are especially important: they rely on the equality or similarity of multiple metadata fields (e.g., identifiers, titles, page numbers) and typically result in a score that indicates the likelihood that two records are duplicates. These candidate duplicates can then be resolved either automatically or after manual checking.

18. Provide guidelines for metadata input or transfer

Improves the accuracy and consistency of metadata

When data collection involves manual input of metadata, it is beneficial to provide guidelines that specify in detail how each metadata category should be recorded. For example, it could specify that publication year for digital publication has to be the year in which the publication became publicly available. Also, it can mention what characteristics of the publication should be used to determine the type of research output. The level of detail depends on the ambiguity in the input of metadata.

Similarly, guidelines can be helpful when metadata collection is carried out by means of metadata transfer from institutional or other national or international databases. Different databases often use different data models and vocabularies, their metadata collection principles might differ from those envisioned in the national database, and other differences can be encountered. For these reasons, it is useful to have guidelines that specify how data should be prepared for metadata transfer.

19. Complete missing data and validate the accuracy of metadata

Ensures that metadata are as complete and accurate as possible

In case of manual input of data or identification of (possibly) erroneous records, the actual research output should be consulted when correcting or completing the record. Even though this is a resource-intensive task that cannot be easily automated, this approach is the best way to ensure the accuracy of metadata. In addition, it helps to avoid replication of erroneous records that are present in other databases.

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.

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.

20. Implement a data validation procedure

Efficiently avoids many data errors associated with (manual) data input

Errors in metadata are often due to typos, misinterpretation of the field content, or diverging conventions for representing a particular kind of content. Automated validation is an approach to identify such errors for data fields that follow a fixed format (e.g., ISSNs consist of 4 digits, followed by a hyphen and 4 additional digits, the last of which can also be an 'X'). It is most efficient to apply data validation during data input: this way, the errors can immediately be corrected prior to being stored in the database.

In addition, one can introduce manual validation to further enhance metadata quality. Assign in each organisation that provides data or participates in data collection one or more experts who are responsible for manual data validation. These persons are tasked with identifying the original research output (in digital or material form) and comparing its characteristics with the database record.

21. Use aggregate statistics to check for systematic data errors

Ensures that no systematic errors propagate in the database

Some errors can have an effect on a large set of records (e.g., all publications from a particular institution) and are as such more serious than errors that only affect individual records. By regularly checking aggregate statistics (e.g., number of publications per year, number of authors per institution, share of each publication type, etc.), many systematic errors can be detected early. Some systems automate this process by, for instance, adding a statistics portal to the regular access to the database.

DATA USE

22. Specify procedures for data access

Enhances the usability of the database

It is useful to develop procedures for data access taking into account the needs of different users and different ways to transfer data. If it is envisioned that data can be downloaded using the user interface of the database, it is important to specify the licence (i.e., which, if any, restrictions apply to the use of the data). Similarly, for an application programming interface (API) and harvesting protocols it is helpful to publicly provide information on how to gain access (if restricted), the terms of use that apply to the use of the system, and the licence that applies to the data.

It is also important to take into account different legal frameworks that might grant access or, in contrast, restrict access to the databases for research output. For example, in some countries, there is legislation like the EU General Data Protection Regulation (GDPR) that protects personal data and legislation that enables access to information that concerns the conduct of public institutions. In this case, while one framework protects data on research output, the other requires their accessibility. A possible solution is to identify what implications follow from different legal frameworks and, in case of overlap, which legislation takes precedence.

23. Offer research output metadata in multiple representations

Ensures that users with different needs and preferences can efficiently use the data

Users of national databases of research outputs use those databases for various needs. Moreover, user profiles and preferences are different. Offering multiple representations of bibliographic records allows users to customize the display and format of downloaded records in accordance with their preferences and needs.

Enable export of metadata using standard, well-known data models – such as CERIF or Dublin Core – and file formats. Support of standard data models makes the integration of data from different systems much more feasible. If possible, implementing the FAIR data principles and usage of Semantic Web file formats like RDF/XML (Resource Description Framework; eXtensible Markup Language) or Terse RDF Triple Language (Turtle) is recommended. The ultimate goal of the Semantic Web is interoperability at the data level without additional implementation performed by software developers.

24. Provide access to the data through a functional user interface

Enables consulting the database in various ways and increases transparency

Databases need a user interface that allows both searching and browsing their contents. The search functionality should both allow for basic search and for specification of advanced, complex, structured queries. Search features should be implemented in accordance with good information retrieval practices (e.g., independence of morphological and inflectional word changes). Browsing should be complemented with the possibility of filtering and sorting the list by publication types, dates and other customisable criteria. It should, in addition, be possible to download selected records or to generate and download a report.

25. Facilitate automated access to the data through an API or a metadata harvesting protocol

Enables automated and efficient use of the database

API access allows external applications to dynamically use (query or download) metadata from the national database. Choose a well-established architecture like Representational state transfer (REST) and a standard format like JavaScript Object Notation (JSON) or XML for the data. Implement authentication and authorisation if usage of the data needs to be restricted.

In addition to API access for generic purposes, access to bibliographic data can also be provided through metadata harvesting protocols for aggregation in other bibliographic systems. The most familiar metadata harvesting protocols are Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), Open Archives Initiative Object Reuse and Exchange (OAI-ORE), and ResourceSync. By implementing the client side of these protocols, it is possible to collect data from local (e.g., institutional) repositories. By implementing the server side, it is possible to export data to bibliographic aggregators at the European or global level, such as OpenAIRE and Networked Digital Library of Theses and Dissertations (NDLTD).

26. Enable crawling of bibliographic records by web search engines

Ensures that database content can be found through regular web search engines as well as academic search engines

Most web traffic is nowadays initiated through a web search engine like Google or Bing. National databases that cannot be indexed by search engines are consequently largely invisible. If broad use of the database by the general public is an explicit or implicit goal, it makes sense to have its contents indexed.

Search engine optimization techniques help to ensure findability. The Robots Exclusion Protocol (https://www.robotstxt.org/) can be used to instruct web crawlers that collect information for web search engines (Google, Bing, etc). Moreover, there are guidelines for specific search engines focused on bibliographic metadata of scholarly publications (e.g., http://scholar.google.com/intl/en/scholar/inclusion.html#overview for Google Scholar).

TRANSPARENCY AND SUSTAINABILITY

27. Encourage feedback from users and other stakeholders

Helps to continuously improve the database in alignment with the needs of users

Create a designated e-mail address or a web form that users can use to provide feedback and suggestions for improvement. This form can also be used to report errors in data and to notify about missing data. Organise regularly live feedback sessions and facilitate meetings of user groups. Feedback can address inaccuracies at the record level (e.g., errors in a bibliographic reference) or at a higher level (e.g., information about a publisher's peer review procedure).

28. Provide up-to-date documentation about the database, its purpose, envisioned uses, limitations, and other aspects

Increases the transparency of the database and its maintenance

Publish documentation (e.g., on a website) to make the database's purpose and envisioned uses explicit. If certain uses are limited or impossible, this should be explicitly stated. Documentation can also help to clarify practical aspects like searching for and downloading specific records. Documentation should be available in all languages that are relevant in the context where the database is operated and preferably also in English. Since only up-to-date documentation is useful, regularly review and update all public documentation.

29. Implement procedures for data provenance

Ensures transparency and openness of the data collection process and contributes to the continuity in data collection

Data provenance at the system level can be provided through documentation of origins of data and methodology and procedures that are used to input, process, and transfer (if applicable) the data. As a general guideline, such documentation should identify every step from the creation (or publication) of an actual research output and the final record in a database. This documentation should be made publicly available and updated.

Data provenance at the record level can be provided through setting up a logging system in the database, which tracks creation, changes, and deletion of individual records, or through annotating records with information on important changes. The latter option is more light-weight but the extent of its usage depends on the staff who input data and maintain the database.

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

Ensures that the database remains up-to-date

Since the context of databases is subject to change, databases need to adapt to ensure that they can continue to fulfil their purposes. Staff responsible for maintaining the database need to deliberately follow research practice, research policy and technological changes and adapt to them where needed. Changes in the database, however, are preferably implemented in combination with data provenance procedures. This increases the long-term usability of the database.

CONTACT INFORMATION ON NATIONAL BIBLIOGRAPHIC DATABASES

Title	Country	URL	Contact person	Contact information
Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB-SHW)	Belgium	https://anet.be/opac/opacvabbg	Raf Guns	ecoom@uantwerpen.be
Croatian Scientific Bibliography (CROSBI)	Croatia	https://www.bib.irb.hr/	Jadranka Stojanovski	jadranka.stojanovski@irb.hr
The Registry of Information about Results (RIV)	Czech Republic	https://www.rvvi.cz/riv	Vendula Kodetová	kodetova.vendula@vlada.cz
The Danish Bibliometric Research Indicator (BFI)	Denmark	https://bfi.fi.dk/	Lotte Faurbæk	lof@ufm.dk
Estonian Research Information System (ETIS)	Estonia	https://www.etis.ee/	-	etis@etag.ee
VIRTA Publication Information Service (VIRTA)	Finland	https://wiki.eduuni.fi/display/cscvirtajtp/VIRTA+in+English	Hanna-Mari Puuska	hanna-mari.puuska@csc.fi

Greek Reference Index for the Social Sciences and the Humanities (GRISSH)	Greece	http://www.grissh.gr/	Irakleitos Souyioultzoglou	irakleitos@ekt.gr
The Hungarian Scientific Bibliography (MTMT)	Hungary	https://www.mtmt.hu/	Andras Holl	andras.holl@konyvtar.mta.hu
Database of Publications in the Social Sciences and Education	Israel	-	Ruti Teitelbaum	szold@szold.org.il
Index to Hebrew Periodicals (IHP)	Israel	http://lib.haifa.ac.il/systems/ihp_eng.html	Neta Waisman	ihp@univ.haifa.ac.il
LOGINMIUR	Italy	-	Marco Mancini	assistenzamiur@cineca.it
Lituanistika	Lithuania	https://www.lituanistikadb.lt/lt Citation data: https://citavimas.lituanistikadb.lt/	Lina Bloveščiūnienė	lina.blovesciuniene@vdu.lt
National Bibliometric Instrument (IBN)	The Republic of Moldova	https://ibn.idsi.md/	Igor Cojocaru	igor.cojocaru@idsi.md
National Academic Research and Collaborations Information System (NARCIS)	Netherlands	https://www.narcis.nl/	Elly Dijk	narcis@dans.knaw.nl
Current Research Information System in Norway (CRIStin)	Norway	http://www.cristin.no/english/	Marit Henningsen	marit.henningsen@unit.no
Polish Scholarly Bibliography (PBN)	Poland	https://pbn-ms.opi.org.pl/	Sebastian Fijałkowski	PBN-HELPDESK@opi.org.pl

Russian Index of Science Citation (RINC / РИНЦ)	Russian Federation	http://elibrary.ru	Gennady O.Eremenko	support@elibrary.ru
The Serbian Citation Index (SCIndeks)	Serbia	http://scindeks.ceon.rs	Nikola Stanić	nikola@ceon.rs
Central registry of publication activity (CREPČ)	Slovakia	http://cms.crepc.sk/	Marta Dušková	marta.duskova@cvtisr.sk
Co-operative online Bibliographic Systems & Services (COBISS)	Slovenia	http://cobiss.si/	Davor Šoštarič	podpora@izum.si
SwePub	Sweden	http://www.swepub.kb.se		libris@kb.se

Source: ENRESSH survey (2017): https://ecoom.uantwerpen.be/sshdatabases

www.enressh.eu