dialogic
innovatie • interactie

Feasibility study Open Knowledge Base – version for consultation

CONSULTATION-VERSION v0.77

This document is a progress report made public for consultation that will lead to a final report.

We welcome feedback on this report, particularly with respect to the following questions:

1. **For whom** is an Open Knowledge Base (OKB) needed (related to support level, see Chapter 2)?
2. Are the identified **core values and use cases** sufficient to render an OKB necessary (see Chapter 2)?
    - If you answer is no: which other (more) important arguments should be considered?
3. Which identified use case(s) should be **prioritized**, and why (see section 2.2)?
4. What **governance model** is preferred? A) One operational team with resources and mandate, B) networked governance with allocated time, resources and mandate (secondment), or C) otherwise (see section 3.1.1)?
    - What is the preferred **role of private parties** in this (e.g., public-private governance)?
5. Should an OKB include **tools and services**? Why (not)? (see section 3.1.7)

Please provide feedback before 13 Jan 2021 to Max Kemman (kemman@dialogic.nl)

Commissioned by:
VSNU

Project:
2020.096

Datum:
Utrecht,
3 December 2020

Auteurs:
dr. Max Kemman
drs. Robbin te Velde

Table of contents

1 Introduction

1.1 Problem statement

Open Science is propelling deep-seated change in the way scientific endeavour is recognised and rewarded. Embedding Open Science demands appropriate infrastructures. Such infrastructures must be sustainable and respect the transparency of interactions between public and private partners in scholarly communications, particularly as the publishing industry increasingly focuses on data analytics services. How data related to publications and other scholarly output is handled and analysed has a crucial impact on judgements about the research success of individual researchers, research institutes, or even national science policies.

Traditionally, universities, university medical centres and other knowledge institutes collect (meta)data related to scholarly communications in institutional systems that are discrete, unconnected, closed and proprietary. As a result, gaining an overview of scholarly communications over multiple institutes is a difficult challenge. Furthermore, in these systems the data is usually closely tied to a particular user interface, which limits the scope of the questions that can be asked and the overviews than can be made.

An open knowledge base (OKB) could address such concerns. Firstly, it proposes that universities start using a **single data infrastructure**. This is very challenging, but there are also strong advantages from building on both the shared human knowledge as well as the technical resources at universities. Secondly, an OKB **separates the graphical user interface from the data**. An OKB opens up (meta)data and allows connecting the underlying (meta)data to other sources of metadata. Such an approach allows for greater freedom – analysis of the data is no longer restricted by the specific way a graphical user interface was designed, nor limited to querying one particular set of (meta)data. The openness also allows third parties to build tailor-made interfaces and additional services on top of the OKB.

Multiple trends within the Netherlands point to the potential of creating an OKB. Policy makers and researcher communities increasingly demand transparency of data and algorithms for responsible decision-making and evaluation; libraries are exploring how infrastructure can offer greater agency in their missions; publishers wish to explore innovative services for fairer metrics for research intelligence and scholarly communication services with high quality Dutch (meta)content; and researchers increasingly expect rapid and trusted access to research outputs and related metrics.

However, despite possible advantages, the idea of an Open Knowledge Base remains ambiguous and needs further elucidation. To further elucidate how an OKB could realistically be embedded within the Dutch research landscape, this feasibility study analyses the different dimensions and considerations underlying an OKB.

1.2 Research questions

The key goal of the feasibility study is to *assess the feasibility of an open knowledge base (OKB) within the context of different options and to make related recommendations pertaining to specific factors such as governance, technical architecture and scope*.

From this research objective, we investigate five research questions that underlie this feasibility study:

1. What are the demands of the different user groups (library IT, national science policy, institutional policy, researchers, private enterprises) with respect to infrastructures containing scholarly communications data?
2. What possible choices can be made in the design of an OKB with regard to the following dimensions?
    a. Governance
    b. Finances and funding
    c. Data scope
    d. Data quality
    e. Service development and commercial engagement
    f. Technical architecture
    g. International context
3. How are dimensions related and what models for an OKB emerge through the combined positions on dimensions?
4. Which model has most support from stakeholders?
5. What are long-term and short-term actions that affect the feasibility of an OKB?

1.3 Method

The results in this progress report are based on interviews with 32 stakeholders. We classified eleven interviewees as Library IT, six as institutional policy, five as national science policy, eight as researchers, and two as private enterprise. See Appendix 1 for an overview of respondents. We have also discussed the concept of an open knowledge base at the UKB Pure User Group and observed the Open Knowledge Base hackathon organised by CWTS and Curtin Open Knowledge Initiative in November 2020.

1.4 Report overview

In Chapter 2 we will analyse the arguments in favour of an OKB from the different user groups. In Chapter 3 we analyse the dimensions and how these lead to three different models for an OKB. In Chapter 4 (missing in this version) we will discuss the timeline for development and implementation of each possible model.

2 Arguments for an Open Knowledge Base

In this chapter we discuss the arguments for an open knowledge base, answering the first research question. In paragraph 2.1 we present the core values underlying an OKB. In paragraph 2.2 we discuss several use cases that may attract engagement and participation from several user groups (national science policy, institutional policy, researchers, library IT).

2.1 The core values of an OKB

The central argument for an OKB lies in the fact that **scholarship is publicly funded** and should, therefore, be publicly available. The public should be able to know what Dutch scholars and scientists work on and what knowledge is produced by Dutch research institutes. The data on scholarly communications, such as publications, are as such arguably **critical information** on scholarship that should be available to the public without any restrictions.

The Dutch government not only finances scholarship, but actively develops policies to foster and sustain scholarship. Article 16 of the 2020 annual budget of the Dutch ministry of Education, Culture and Science states that the ministry is tasked with *financing, stimulating and directing* Dutch scholarship to create and sustain an internationally competitive research environment.[1] The Dutch government subsequently regularly sets agendas prioritising specific research fields at the expense of others, for examples in programmes such as the Dutch Research Agenda[2] or Topsectoren[3]. Approximately two-thirds (€655 million) of the annual budget of the national science funder NWO is allocated for thematic research.[4] Yet despite the clear political importance of such allocations, **gaining a comprehensive and publicly available overview of the results from policies and allocations is currently not feasible**. Assessments of scholarly activity today largely depend on paid services such as Web of Science (Clarivate) or Scopus (Elsevier).[5] Furthermore, these services predominantly index journal articles and do not adequately cover other scholarly output such as books, code, data, et cetera.[6] Finally, these services no longer truly fit the Dutch commitment to alternative assessments of Dutch scholarship.[7]

---

[1] Rijksoverheid (2020). *3.12 Art. 16. Onderzoek en wetenschapsbeleid*. Rijksbegroting [rijksbegroting.nl], consulted 26 November 2020.

[2] [nwo.nl]

[3] [topsectoren.nl]

[4] KNAW (2019). 'Evenwicht in het wetenschapssysteem. De verhouding tussen ongebonden en strategisch onderzoek.' Amsterdam: KNAW.

[5] Any evaluation of a specific research field, research organisation or thematic domain requires the purchase or licensing of proprietary bibliometric data from one or more of the large publishers. In concrete terms, these amounts range from €10.000 to over €100.000 per study.

[6] Jeroen Bosman and Bianca Kramer (2019). 'Publication Cultures and Dutch Research Output: A Quantitative Assessment'. Zenodo.

[7] San Francisco Declaration on Research Assessment [sfdora.org]; NWO (2019) KNAW, NWO, ZonMw to sign DORA declaration. [nwo.nl]; VSNU, NFU, KNAW, NWO, ZonMw (2019). Ruimte voor ieders talent: naar een nieuwe balans in het erkennen en waarderen van wetenschappers. Position paper;

Moreover, the data on scholarly communications is critical for the reports that institutes are obligated to produce, notably the SEP (Standard Evaluation Protocol) and KUOZ (Key figures on University Research)[8] reports. A limitation of the **SEP and KUOZ reports** with the current situation is that institutes report aggregated data, while the data remains in closed systems. Yet several interviewees stated that the data is incomparable between institutes as a result of differing definitions and interpretations, e.g. on what counts as a publication. As a consequence, aggregated data is incomparable between institutes and of little value at the national level (e.g., for policy purposes). Furthermore, evaluations based on specific **metrics cannot be verified** afterwards due to lack of access to the underlying data. This is in sharp contrast to the Leiden Manifesto[9] that argues for transparent and verifiable research metrics. However, interviewees noted that disclosing how these metrics are calculated is inherently at odds with existing business models built around the proprietary nature of these metrics and algorithms. In contrast, an OKB could offer **transparent data as well as algorithms** to render indicators verifiable.

Finally, there are good reasons to launch the OKB at this very moment. First, in the private sector, large publishing companies are increasingly moving to service-oriented models based on data ('Platform economy')[10]. As a result, there is an increasingly unlevel playing field between research institutes and large publishing companies with respect to data access. To prevent **data lock-in** and vendor lock-in, there is currently opportunity to move data to an open infrastructure. One interviewee argued that an OKB in that sense represents an "exit-strategy" for universities in case they do not extend current contracts with large publishing companies. Second, in the public sector, discussions related to open science and open access have gained momentum in recent years. For instance, the current way in which academic output is measured is increasingly criticized, such as debates on the recognition of academic impact 'Erkennen & Waarderen'.[11] The need for new evaluation metrics of scholarship likely requires rethinking of underlying **infrastructure for assessments** of scholarly communications as well. Finally, the current system of financing scholarship is increasingly under debate as well, calling for instruments that complement the aforementioned allocations for thematic research. Recent reports by the KNAW have argued in favour of rolling grants to sustain continuous funding for innovative fundamental and applied research.[12] An OKB could facilitate future **evaluations of the different methods of funding**.

2.2 Use cases for an OKB

While the above core values of an OKB may already be sufficient for agreement and engagement with the *concept* of an OKB, engagement with **actual implementations** of an OKB depends on what **use cases** are supported. In this section, we discuss several use cases identified in interviews for four possible user groups: national science policy, institutional policy, researchers, and library IT.

---

[8] VSNU (2019). Definitieafspraken Wetenschappelijk Onderzoek: Toelichting bij KUOZ.

[9] Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429-431.

[10] Aspesi, C., & Brand, A. (2020). In pursuit of open science, open access is not enough. *Science*, *368*(6491), 574-577; Schonfeld, R. C. (2017). When is a Publisher not a Publisher? Cobbling Together the Pieces to Build a Workflow Business. *Scholarly Kitchen*. [scholarlykitchen.sspnet.org], consulted 5 November 2020.

[11] VSNU, NFU, KNAW, NWO, ZonMw (2019). Ruimte voor ieders talent.

[12] KNAW. (2020). *Het Rolling-grantfonds—Kloppend hart voor ongebonden onderzoek*. KNAW.

2.2.1 National science policy

For national science policy, the core values described above were regularly identified as of sufficient importance. Additionally, a use case that interviewees identified is that science **funders currently have a limited overview of the scholarship that is produced** as a result of funding grants. In the current situation, scholars (manually) enter their publications both in their university CRIS system as well as in the NWO ISAAC system[13]. One metric that funders want to assess is the number of publications that are open access; in principle, 100% of all publications supported by public funding should be open access. However, since the information is entered manually into ISAAC, the data contains errors. Researchers and institutes furthermore have differing definitions of what counts as a publication coming from a grant, where these differing definitions are not made explicitly but hidden in closed systems.

Metrics that are commonly used today such as *h*-index, journal impact factor (JIF) or number of publications are increasingly critiqued for their limited and biased view of what scholarship should achieve. Current indicators of scientific quality can lead to perverted incentives to score highly on specific metrics. The current system is furthermore noted to lead to vicious cycles of a small elite of researchers being able to attract the majority of funding, while the majority of scholars are left struggling ('Matthew effect').[14] For this reason, scholars and policy makers are increasingly considering the adoption of so-called **next generation metrics** which aim to measure what matters for scholarship and society. For example, science policy makers might wish to monitor how Dutch scholarship contributes to the United Nation's sustainable development goals (SDG).[15] For the Netherlands, metrics of national interest may monitor how Dutch scholarship contributes to mission-oriented research and key enabling technologies.[16]

The European Commission expert group on Altmetrics argued in its report that "[n]ext generation metrics should be underpinned by an open, transparent and linked data infrastructure".[17] Such metrics should moreover not be used in a singular fashion (one indicator to rule them all). The European Commission working group on rewards under open science instead emphasised the need for multi-dimensional criteria, using metrics that are appropriate and relevant by tailoring to individual researchers.[18] Finally, recent discussions

[13] [nwo.nl]

[14] José van Dijck en Wim van Saarloos (2017). 'Wetenschap in Nederland: waar een klein land groot in is en moet blijven' Amsterdam: KNAW.

[15] For example, see Armitage, C. S., Lorenz, M., & Mikki, S. (2020). Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results?. *Quantitative Science Studies*, *1*(3), 1092-1108; Aurora Universities Network (2020). SDG Analysis: Bibliometrics of relevance; VSNU (2019) SDG-Dashboard: Impact Nederlandse universiteiten in kaart gebracht, [vsnu.nl], consulted 26 November 2020.

[16] Ministerie van Economische Zaken en Klimaat (2019). Missies voor het topsectoren- en innovatiebeleid.

[17] European Commission. Directorate General for Research and Innovation (2017). *Next-Generation Metrics: Responsible Metrics and Evaluation for Open Science.* LU: Publications Office, p. 15.

[18] European Commission. Directorate General for Research and Innovation (2017). *Evaluation of Research Careers Fully Acknowledging Open Science Practices: Rewards, Incentives and/or Recognition for Researchers Practicing Open Science.* LU: Publications Office. [doi:10.2777/75255].

have pointed to the need not to use metrics for benchmarking and ranking, but instead for providing the means to evaluate institutional or national strategies or societal agendas.[19]

An OKB could provide the infrastructure for **more reliable and relevant (next generation) metrics that better align with national priorities**, where it is transparent what data underlies those metrics. An additional advantage could be that researchers no longer have to enter data both into their institutional CRIS as well as into ISAAC, since the data from these systems is connected.[20]

2.2.2 Institutional policy

Interviewees from institutional policy did not unanimously see use cases for an OKB. Interviewees noted that institutional researchers and business intelligence were not the central stakeholders in the decision to adopt Pure CRIS systems. One interviewee argued that universities generally do not set strategies for publications but for populations of students and staff. Data on HR and student populations is generally well available and provides immediate opportunities for strategy setting. However, this interviewee did note that the lack of interest in scholarly communications data may simply be a consequence the current absence of accessible high-quality data. When envisioning what an OKB could support, they did see opportunities for better informed **benchmarking of research between institutes**, discovering which institute is best positioned or quickly gaining a better position for specific research topics and the setting strategies for their institute to visibly position themselves on research topics. Furthermore, university policy makers agreed that **SEP and KUOZ reports** are obligated reports where an OKB could prove useful (i.e., greatly facilitate the drafting of the reports, which is nowadays still largely done by hand, and thus a labour-intensive and hence costly and error-prone process).

Additionally, research manager and institutional policy makers can benefit from the next generation metrics discussed in the context of national science policy above. Such metrics might facilitate the **analysis of institutional strategies** with regard to SDGs or mission-oriented research and key enabling technologies. Furthermore, institutional managers could develop their own metrics and assess how their institute compares to other institutes depending on what they find strategically important.[21]

2.2.3 Researchers

Since the data in an OKB is on scholarly communications, some interviewees noted that researchers themselves are generally not the target user group, with the exception of some researchers in bibliometric analysis. The advantage an OKB could provide is in **supporting researchers to track their output**, for example by connecting CRIS and funding systems so they do not have to enter data twice. Some researchers saw opportunities for services that could enable them to better keep track of who is working on similar research topics.

---

[19] Ingrid Bauer et al. (2020). 'Next Generation Metrics'. [doi:10.5281/ZENODO.3874801]; Elizabeth Gadd (2020). 'University rankings need a rethink'. *Nature* 587, nr. 523.

[20] Note that the IT-principle of single point of data entry has been transposed into a legal obligation in the policy domains of income and labour ('Wet eenmalige gegevensuitvraag werk en inkomen' [wetten.overheid.nl]).

[21] Elizabeth Gadd (2020). 'University rankings need a rethink'.

A second use case is to improve the public **exposure** of researchers' work. For instance, specific metrics could be constructed on top of the OKB that makes the societal impact of the scientists' efforts (better) visible.

2.2.4 Library IT

Finally, an important user group is library IT, who are largely responsible for the current CRIS systems and will likely become responsible for providing the data or facilities for an OKB. Interviewees noted that one important case is that **CRIS systems across institutions contain a large amount of duplicate data**. Universities keep track of collaborating partners, which can run into thousands of organisations (for example, other universities or companies). Tracking this information in a shared OKB could save university librarians significant amounts of work. Sharing information about organisations could furthermore increase data quality, since enrichments to data are made available to all universities. A central concern for library IT is then to what extent the data of an OKB could be fed back into local repositories and/or CRIS systems. Taking this concern one step further, a question for library IT is to what extent an OKB could eventually replace their current CRIS systems.

3 Long-term prospects of an Open Knowledge Base

In this chapter we elaborate what a Dutch open knowledge base could entail and what model has most support from stakeholders. We thereby answer research questions 2, 3 and 4. In paragraph 3.1 we first discuss the different dimensions and what choices can be made with respect to each dimension. In paragraph 3.2 we identify three models (API-standards, Warehouse, Research Environment) that emerge through the combinations of positions on the different dimensions. In paragraph 3.3 we discuss which model has most support from stakeholders.

3.1 Dimensions underlying an OKB

3.1.1 Governance

Main considerations for governance

- Interviewees agreed governance should be completely public
- Top-down or bottom-up (operational team employed at a single legal entity; a network of experts; a combination of both)
- Pace of development was emphasised to prevent commercial actors from surpassing an OKB as well as preventing local actors from undermining a consortium approach
- Governance requires approximately five to fifteen people, with expertise related to project management, data management, data architecture, legal affairs, account management, and possible software development and user experience

The first and perhaps most fundamental dimension concerns who should be in control of an OKB, notably whether this should be **public or private**. The current situation where universities individually license commercial software in this respect presents a model where governance is entirely commercial. Furthermore, at the moment the vast majority of Dutch institutes have opted to use Pure as their CRIS system.[22] Consequently, at present governance lies with Elsevier as developer, maintainer and owner of the CRIS software.

Interviewees agreed that while development of an OKB could eventually be done commercially, the **governance should be completely public**. This entails that a commercial software developer licenses the software to be owned by a public organisation or that the software is licensed to be open source. Interviewees emphasised this final model, which would make an Open Knowledge Base also open in itself as a system that can be replicated, adapted and distributed.

Interviewees did not consider public-private governance, where software is owned and governed by both public and private parties, to be a desirable situation. The main reason for this was that this still risks vendor lock-in, since a private party has commercial interests to enable, disable or veto specific features of an OKB that may be relevant to the Dutch public sector. However, in one interview the option was raised to participate in existing non-profit infrastructure projects, rather than initiate a new one. In this scenario, the Netherlands would participate in existing international infrastructures by licensing existing systems or participating in international consortia. While this could limit some features desired by the

[22] [elsevier.com]

Dutch public sector, this was argued to represent a stronger case since it starts from the international connection (see below).

If, however, an OKB were to be governed by a Dutch public consortium or party, another aspect is the extent to which this should follow a **top-down or bottom-up governance**. In a bottom-up (or grassroots) approach, the Dutch research institutes establish a community to discuss and agree together on decisions with respect to an OKB. An advantage of this model is that institutes are engaged democratically and develop mutual trust and respect through coordination. An OKB is then a shared project that the participating institutes all recognise and desire. A major downside of a bottom-up approach may be that institutes lack incentive to agree and act and that progress stalls. When progress stalls, individual institutes may find it necessary to act on their own, undermining a collective (consortium) approach.[23] Furthermore, multiple interviewees noted that research institutes and other actors in the Dutch academic landscape **currently lack the necessary expertise to sustain an OKB** (see below). Creating roles and responsibilities that require new personnel may be more efficient when centralised in a top-down approach. Several interviewees boldly stated that it is vital for a viable OKB to have a legal entity on its own (see below).

In a top-down approach, a small consortium of actors mutually agrees on decisions and set this as the agenda for the Dutch research community. Actors that could participate in such a top-down consortium that were mentioned in interviews included SURF, VSNU, DANS, but also stakeholders such as NWO, KNAW and NFU. The advantage of a top-down model is that governance is centralised; decisions can be made much faster and easier and actions can be initiated from the centre. Several interviewees noted the **importance of pace** of development to prevent being surpassed by commercial actors which can quickly provide working systems that are however not open. A limitation of existing systems such as NARCIS and OpenAIRE that was mentioned in interviews is that they can only collect the data that is provided by research institutes and have no mandate to provide feedback on the scope or quality of that data. As a result, several interviewees noted that the data in these systems lacks in quality and therefore has limited practical value. An opportunity for a centralised governance model on an OKB could be to provide a **mandate to request or demand better data** to ensure high data quality and utility (see also §3.1.5 below).

Based on interviews and comparisons with other systems, we estimate that the governance team should consist of approximately five to fifteen people, depending on the model chosen for an OKB. At the very least, this team should include roles and responsibilities related to:

- Project management (both during development and when operational)
- Data management
- Data architecture (e.g., Linked Data)
- Legal affairs (at least copyright and intellectual property, possible GDPR and privacy)
- Account management (including support)

A consequence of a top-down model is that the central organisation needs to establish a process for **account management**, providing support to research institutes and acquiring insights and feedback. This account management requires a sound legal organisation within which a fairly stable group of people operate that are recognisable as concerned with and

[23] This risk finds precedence in the failure of a Dutch consortium led by SURF to agree on a uniform CRIS system, after which Dutch institutes each individually had to license CRIS systems; most ended up choosing Pure.

responsible for an OKB both on a day to day and strategic basis. This has financial consequences for an OKB, see below.

Depending on the model chosen for an OKB and whether an OKB should depend on external (commercial) parties for the development of tools and services (see §3.1.7 below), the above roles may need to be extended with the following:

- Software development
- User experience/human computer interaction (to ensure the usability of tools and services for different user groups)

A risk of a top-down approach, however, is that research institutes lack incentive and trust in the process. Institutes will then not engage but conform to the very minimum of what is demanded.

A **middle ground may be to organise networked governance** consisting of a central task force with experts employed at research institutes. These experts should then receive an official mandate to make central decisions and should receive allocated time (up to full-time) to work on an OKB.[24] Depending on the model chosen for an OKB (see section 3.2 below), configurations are possible where a central party maintains a system, receiving input from working groups consisting of experts employed at research institutes.

3.1.2 Critical mass

Main considerations for critical mass

- Interviewees agreed a subset of research institutes would be sufficient to start rather than all research institutes at once.
- It is important to make the benefits for early adopters visible.

The key question with regard to critical mass is whether an OKB needs participation from *all* Dutch institutes from the start, or whether this can grow over time. Interviewees agreed that **an OKB does not need to start with all institutes**, but that a critical mass of early adopters can be sufficient. A target here may be at least five institutes, which was mentioned in one interview as a general rule of thumb for SURF consortia. In this case it is, however, important to **make it clear what the benefits are for early adopters** and also who covers the start-up costs. An additional remark is that an OKB that is based in existing CRIS systems should relatively easily gain critical mass, since the Pure CRIS system already presents a critical mass of institutes. Obviously, this gives a big lead to the owner of Pure, Elsevier.

3.1.3 Technical architecture

Main considerations for technical architecture

- Federated or centralised (interviewees noted the opportunities for centralised architecture)

The central question for the technical architecture is whether to follow a **federated or centralized approach**. In a federated approach, an OKB is a connected infrastructure of

[24] An example of such a networked yet centralized organization is the Netwerk Digitaal Erfgoed (Network Digital Heritage), which consists of experts employed at cultural heritage institutes who have allocated time, resources and a mandate to work on this network [netwerkdigitaalerfgoed.nl]. Another example of such a network is NOVA (Netherlands Research School for Astronomy) [nova-astronomy.nl]

systems running locally at the individual participating institutes. An advantage of this approach is that institutes remain in control of their own progress; some institutes may connect sooner with an OKB, while others follow later on. Furthermore, institutes remain in control of their own data; institutes can define access policies to sensitive data where necessary and can disclose data at their own pace. Institutes can provide the minimally requested data but can in principle decide to disclose more data than demanded, providing further enrichments. A federated model thus has the advantage of **organizational scalability** in setting up an OKB. By simply following a set of standards and disclosing their data in the appropriate way, institutes themselves can initiate participation in an OKB.

However, a major disadvantage of a federated approach is the lack of **technical scalability**. Querying or analysing the data requires users either to download data dumps from all other participating institutes or requires an infrastructure to approach the data at each individual institute. Several interviewees argued that this is not satisfactorily for users of the data, since they either run into limitations of how many queries can be sent to each individual system (one interviewee noted it could take several months to request all the necessary data from all Dutch institutes). Furthermore, analysis is limited to the sustainability and performance of each individual institute; one interviewee with several decades of experience in scalable infrastructure noted they had not yet seen any good example where a federated approach scaled well beyond five or six institutes. Finally, several interviewees argued that institutes often lack technological expertise for implementing and sustaining advanced systems. For example, if an OKB were to be designed as Linked Open Data (LOD)[25] then implementation and sustainability is hampered by the lack of expertise with LOD approaches at the institutional level.

In a centralized architecture all data is instead collected in a single, centralized, system. Analysis is easily scalable, since all data is accessible from a single point, in contrast with a federated architecture. Another advantage of a centralized system that was mentioned in several interviews is that data quality can be more easily harmonized; there is a central overview of possible gaps and inequalities between data that can be used to provide feedback to institutes providing the data, or that may be enriched through other means (algorithmic or manual curation). Furthermore, data can be enriched by collecting data from other systems than the institutional CRIS systems, for example adding data from other open infrastructures such as Open Citations, Crossref, ORCID, or commercial providers such as Microsoft Academic Search (MAS).[26] A possible disadvantage is that the development of a centralized architecture is less scalable, since it is more difficult to add different data fields later on. There is then an increased risk of path-dependency, where design choices early in the process determine possible research questions in the future.[27] Finally, a disadvantage is that the costs of initial development have to be made in full before the first institute can add data. This then creates a risk of uneven costs between institutes participating early on and those joining later. Finally, a risk is that this centralized architecture is merely used for depositing data, but eventually ends up not being used for research, as some interviewees critically argued is the case with NARCIS. Note that for the end user the infrastructure (OKB) is not so much important, but rather the quality of the data (content) and the user-friendliness of the services that run on top of the infrastructure.

---

[25] [lod-cloud.net]

[26] [ma-graph.org]

[27] This risk can partially be alleviated by considering the centralized warehouse as part of a network including other (institutional) databases which may provide more flexibility.

3.1.4 Data scope

Main considerations for data scope

- Most interviewees saw metadata as most feasible, rather than only identifiers or all content including full texts
  - Abstracts would be desirable to be included as part of metadata, but depends on copyright
- Interviewees noted the opportunity for an OKB to emphasise non-traditional research output in contrast with existing bibliometric systems
- Fine-grained personal data facilitates use cases but adds complexities related to privacy and GDPR

Interviewees agreed that an OKB should at the very least contain data with respect to publications and grants, as well as producing entities such as authors, institutes and other affiliations, and funding agencies. Interviewees disagreed however about how much data should be collected about publications and grants; **only identifiers, metadata,** or **full texts**.

In the minimal model, an OKB contains only identifiers and relations between those identifiers. An OKB then contains a list of publications in the form of digital object identifiers (DOIs), Handle identifiers, or others. These publications are linked to authors, who are represented by a list of ORCID identifiers. Authors are subsequently linked to their institutional affiliation in the form of ROR or ISNI identifiers. An OKB then functions similar to the "yellow pages" in providing identifiers that may be used to retrieve more data from other services such as Crossref for publications or ORCID for authors. Most **interviewees were not in favour of this minimal data model**, since it provides very little opportunity for analysis but puts the burden on users to collect data from other systems.

In a more encompassing model, an OKB contains the metadata related to publications, grants, authors and organisations. An OKB then contains the identifiers from the minimal model but extends this with additional metadata such as title, publication venue, year, names, locations, et cetera. In principle this is the data that is contained in CRIS systems and allows for analyses on the output of institutes, research groups and individual researchers. For example, an Open Access monitor, one of the use cases identified in the previous chapter, would be facilitated well by the metadata model. **Most interviewees saw the metadata model as the most feasible**. In principle, metadata in CRIS systems is owned by the institutes and can be made publicly available without copyright restrictions.[28]

An open question is to what extent this metadata model could contain **abstracts**. Abstracts could provide the means for several services such as an SDG classifier (see §2.2.1 above). However, abstracts are not clearly part of open metadata. The Initiative for Open Abstracts estimates that 6.6% of all works with a Crossref DOI and 8.3% of journal articles disclosed their abstracts via Crossref.[29] When abstracts are available via Crossref they can be

[28] For example, NARCIS places no restrictions on reuse of metadata which is aggregated from institutional repositories (with the exception of metadata on persons and organisations) [narcis.nl]. Metadata disclosed via Crossref can be reused without restrictions as well [crossref.org]. An exception to this openness of metadata may be when individual institutes have transferred ownership of metadata to private parties in institutional CRIS license contracts.

[29] Initiative for Open Abstracts (2020). [i4oa.org], consulted 26 November 20.

considered public and open data.[30] One interviewee noted that under the Taverne Amendment, abstracts are openly available via institutional repositories.[31] An OKB could then at least provide identifiers as part of metadata to retrieve abstracts from institutional repositories.

Finally, in the most complete model, an OKB contains the identifiers and metadata as well as the full texts of publications and possibly (successful) grant applications. This allows much more advanced analyses such as text mining to detect emerging topics or analysing contributions of researchers to specific topics. Full texts are however not always openly available. In this case an OKB could include identifiers to retrieve the full texts available at institutional repositories under the Taverne Amendment. Whether full texts should be included furthermore depends strongly on the use cases identified; while it enables the most advanced research question, it adds complexities both on a technical level (such as whether an OKB should accept DOCX, PDF, DOT or other files) as well as on a legal level. Since this model introduces additional complexities, several interviewees noted full texts could be added later on, rather than be included from the start. It should be noted that here, too, Elsevier has a big lead since it already has a large number of full texts at its disposal.

A second consideration with regard to data scope is with respect to **non-traditional research output** such as datasets, software or scholarly communications aimed at the public. Such data should be included to align with next generation 'Erkennen & Waarderen' metrics (see §2.2.1 above). Furthermore, an OKB should adequately cover the research output from disciplines that are not served well by current bibliometric systems such as the humanities and social sciences which publish in books, journals with DOI's and in Dutch or other languages.[32] While such metadata may be more difficult to retrieve and collect compared to traditional research output with DOI's, several interviewees as well as the participants of the Open Knowledge Base hackathon noted that **this is necessary to create an advantage to existing bibliometric systems**, rather than repeating the same biases.

Finally, a third consideration is the extent to which an OKB should contain data on scholars themselves. For example, while a paper usually only mentions the institutional affiliation on the university level, **more fine-grained HR data** may provide insight into affiliations related to specific research groups or to what extent scholars work on temporary contracts tied to specific project funding (which in turn allows linking a publication to a grant). Personnel data may furthermore assist institutional assessment on equality, diversity and inclusion (EDI) related to sex, ethnic background and age.[33] This may however introduce complexities (and opposition) in making such fine-grained **personal data** publicly available. With respect to governance including such data necessitates the addition of a privacy officer and/or GDPR legal expert.

3.1.5 Data quality

Main considerations for data quality
- Data quality deemed essential
- Improving data quality can occur at institutional and/or central level

---

[30] European Commission. Directorate General for Research and Innovation. (2019). *Open metadata of scholarly publications: Open science monitor case study.* Publications Office. [doi:10.2777/132318]

[31] You share, we take care! (n.d.) [openaccess.nl], consulted 26 November 2020.

[32] Jeroen Bosman and Bianca Kramer (2019). Publication Cultures and Dutch Research Output.

[33] Curry, S. et al. (2020). *The changing role of funders in responsible research assessment: Progress, obstacles and the way ahead.* Research on Research Institute. [doi:10.6084/M9.FIGSHARE.13227914]

With respect to data quality, interviewees underscored the problem that no data source is complete or perfect. However, an OKB consisting of low-quality data **risks the utility of an OKB**, with the risk that stakeholders end up not using an OKB. Several interviewees noted that low-quality data is a reason for the limited usage of NARCIS. For example, if not all institutes provide data on the open access status of publications equally, a national Open Access Monitor cannot be sustained using an OKB. Interviewees agreed that an advantage of basing an OKB in the data from institutional CRIS systems is that these are relatively complete, since researchers are required to list all their publications, and that affiliations are fairly easily made, since institutes know where their researchers work. A disadvantage is, however, that there are large differences in quality between institutes, between research groups or even between individual researchers. Since the data is provided (partially) manually, fields are not entered or filled with errors.

A question that interviewees raised is whether data should be cleaned up at the institutional level (providing feedback to institutes to clean up their data) or at the central level. Depending on the governance model chosen (see §3.1.1 above), additional requests or demands could be returned to research institutes to improve the data provided. At the central level, an OKB could provide community curation options or enrich data automatically from other sources such as Crossref, Microsoft Academic, ORCID and others. One interviewee raised the possibility of outsourcing data quality control for manual curation (e.g., in India). All interviewees agreed, however, that effort should be put in an overall increase in data quality as much as possible. Such enrichments could potentially be fed back to institutional CRIS systems for local data improvements, depending on how CRIS systems can retrieve and import this data. Furthermore, at the OKB hackathon participants agreed that **clear inclusion of provenance** of data is necessary to sustain data quality.

3.1.6 International connection

Main considerations for international connection

- National or international OKB (interviewees agreed national OKB is feasible)
- National OKB should follow international data standards
- National OKB may lack certain use cases

With respect to the dimension of international connection, the question was whether the Netherlands can develop an OKB on a national scale, or whether an OKB would need to engage international participation from the start. **Most interviewees agreed that the Netherlands can first develop an OKB on a national scale**. Several interviewees warned that seeking international cooperation from the start would risk turning an OKB into a slow and nearly impossible negotiation process. Interviewees argued that the Netherlands has a sufficient critical mass of research institutes that operate in a level-playing field and of high excellence; a Dutch OKB can then be fairly easily negotiated among Dutch stakeholders, while setting an example for other countries to follow.

Interviewees, however, did underscore the need to **follow international standards** as much as possible, rather than defining a new set of standards without international alignment. One interviewee gave the example of the digital author identifier (DAI)[34], a Dutch standard for assigning authors with an identifier that failed to gain traction outside of the Netherlands. While the DAI is still used in for example NARCIS, most other infrastructures

---

[34] [wiki.surfnet.nl]

nowadays use ORCID identifiers. By following international standards, a Dutch OKB could in the future be connected to international infrastructures, rather than risk isolation.

A possible **limitation of starting at a national scale may be in the use cases that are facilitated** by an OKB. A national OKB facilitates comparisons or benchmarking between Dutch institutes but does not enable benchmarking with institutes from other countries. Assessments of the national quality of scholarship may require international data. Furthermore, the Dutch academic ecosystem is strongly international, characterised by international collaboration as well as international mobility. While a nationally oriented OKB contains data on co-authorship and may thereby include insights into international collaboration, it will likely miss publications from Dutch researchers for periods in which they work abroad. When an OKB underlies use cases for individual assessment of scholars, such gaps need to be recognised and/or accounted for.

3.1.7 Service development and commercial engagement

Main considerations for service development and commercial engagement

- Interviewees agreed development may be done by commercial parties
- Essential to separate development of the data infrastructure and of the tools and services
- Interviewees disagreed on the inclusion of services and tools (several interviewees argued an OKB needs basic tools to cover and demonstrate main use cases)
- Several interviewees noted a Share Alike data license would be desirable to prevent data lock-in of enrichments

An OKB is essentially a data infrastructure on top of which services can be developed. **Interviewees disagreed however whether services should be part of an OKB** or left entirely outside the scope of an OKB as a data layer. As noted in the introduction, one of the reasons for an OKB is to separate the graphical user interface from the data.

Some interviewees argued that an OKB should only be the data layer on top of which users, institutes and private parties can develop services. However, other interviewees were critical that this puts the burden on users to develop tools and services, leading to the risk that ultimately the data sits unused (see before, §3.1.3). The use cases described in chapter 2 relate to **services**, not to data repositories per se. If an OKB were to engage users, it should provide services that users care about (e.g., people care about the trains that get them from place A to place B, not about the rails). They **argued that an OKB should at least provide basic tools that cover the main use cases for an OKB**. Some interviewees argued for even more advanced tools such as virtual research environments (VRE) or dashboards that provide a single point of access for overviews and analysis.

Interviewees did agree that service development should engage private parties. Several interviewees noted that there is a lack of expertise at both the institutional as well as at the national level to develop such tools and services. When hiring private parties, interviewees argued that this should concern commercial services such as software development or data curation after which the product is publicly owned. That is, when engaging a commercial enterprise, the resulting software should be open source or data enrichments should be open data. It is moreover **essential to separate the development of the data infrastructure and of the tools and services** to prevent vendor lock-in, which however does not necessarily exclude the possibility of both being conducted by the same (commercial) party. An OKB should, however, not prevent commercial enterprises from developing closed systems on top of an OKB for commercial gain. Finally, one interviewee argued that service development and innovation should not be left entirely to private parties, since this situation risks bringing about new commercial dependencies and vendor lock-in. They argued that public governance should include a minimum team of developers for continued development and innovation (if only to ensure 'absorptive capacity'[35] in the public sector).

One question for an OKB is what kind of data license would be most appropriate to fulfil the above characteristics. To make the data completely public and useful for both public and private parties, a CC0 license[36] (used by for instance Wikidata) could be considered that provides opportunities for both public and commercial services. However, to **prevent data enrichments to end up in closed systems and new forms of data lock-in**, a CC BY-SA license[37] (used by for instance Wikipedia) could be considered that allows commercial usage of the data as long as the data enrichments are shared under the same license.

3.1.8 Finances and funding

Main considerations for finances and funding

- The costs of an OKB strongly depend on the model chosen
- A centralised warehouse may cost €1-2M in start-up costs and €0.5-1M in annual operational costs, excluding the efforts required from research institutes to provide data
- Governance should prevent early adopters from bearing disproportionate weight in funding during start-up

Central questions here are what an OKB would cost and who should cover these costs. This dimension largely depends on choices made with respect to the dimensions discussed above. Several interviewees stressed that too much focus on the financial aspect risks an OKB starting as a cost-savings exercise rather than from intrinsic arguments for an OKB related to core values and use cases.

While an exact provision of costs is beyond the scope of this feasibility study, based on experiences from other scholarly infrastructures (and our own IT experience) realistic estimates of the order of magnitude can be made. First, with respect to the technical aspect, this depends to some extent to the scope of data (see above) as well as the technical infrastructure.

In case a central database is developed, the start-up costs were estimated in two interviews at roughly one to two million euros (software development and acquisition of hardware or licensing of cloud services). The costs of on-going development and innovation largely depends on the scope of innovation and service (discussed above) but was estimated to be between half a million to one million euro annually in personnel costs. Maintenance costs were estimated at roughly one million euros per year (personnel and hardware or cloud licenses). Maintenance costs are mainly related to personnel necessary for account management. For an overview of necessary personnel, see §3.1.1. Furthermore, the participating research institutes need to invest personnel in training, usage and data provision, which could cost several million (mostly in-kind) both during start-up as well as in on-going costs.

---

[35] Cohen and Levinthal (1990), "Absorptive capacity: A new perspective on learning and innovation", Administrative Science Quarterly, Volume 35, Issue 1 pg. 128-152.

[36] [creativecommons.org]

[37] [creativecommons.org]

To **prevent early adopters from bearing disproportionate weight in funding**, start-up costs should be covered by a stand-alone, one-off funding, for example from the Ministry of OCW (Education, Culture and Science) or national science funders. Recurrent costs for maintenance could subsequently be covered by annual fees from participating institutes, for which several models are available (fees could for instance be based on the size of the institute and/or the volume of actual use).

3.2 Possible OKB models

In analysing the above dimensions, we arrive at three possible models for an OKB and suggest how these are positioned on each dimension. Below we introduce the three models.[38]

The first model is what we call the **Open Knowledge API-standards**. In this model, an OKB is merely a set of standards of metadata that each institute or organisation should provide through an openly available API. Governance is *public-private*, insofar as standards can be agreed through public governance, but APIs on CRIS systems are developed and controlled by private parties. The technical architecture is *federated*, where each institute or organisation is responsible for their own API endpoint. An advantage of this model is that API-standards can grow over time, for example first for publications and later for grants. At the same time, once an API-standard is agreed and implemented, an OKB quickly grows substantially since many institutes can implement the same standard at the same time. For example, an API for CRIS systems could relatively quickly gain critical mass since the majority of Dutch institutes use the same CRIS system (Pure). A risk of this model is, however, that is ends up not truly open in the sense of an API without limitations to read, mix and share data.

The second model is what we call the **Open Knowledge Warehouse**. In this model, an OKB is a tangible database or network of interconnected databases in which data is stored within the scope of the OKB. Governance is *public*, insofar as the warehouse is in control by public parties, but possibly developed by commercial parties who agree to public or open source and open data licenses. The technical architecture is most likely (but not necessarily) *centralized*, which provides the advantage to harmonize data quality and to centralize the necessary expertise for development and maintenance. An advantage of this model is that the data is stored in an open system and is available through a single point of access, in contrast with the API-standards where data is stored in closed systems and accessible through a multitude of API endpoints. A disadvantage of the Warehouse model is that costs are increased compared the API-standards model. Furthermore, there is an increased risk of path-dependency, where design choices impact possible research questions in the future, with more difficulty to expand the system at a later stage. However, this may be alleviated by the advantage that a centralized Warehouse could better integrate with existing infrastructures such as Microsoft Academic Search (MAS), Crossref, ORCID or DataCite, enabling further enrichments of the data without additional demands of CRIS systems.

If an Open Knowledge Warehouse is to be based on CRIS data, this model should be understood as an extension of the API-standards model rather than a replacement. To deposit data in the Warehouse, API-standards will be necessary to extract data from the institutional CRIS systems.

The third model is what we call the **Open Knowledge Research Environment**. In this model, an OKB is a research environment (usually a Virtual Research Environment – VRE) in which the data within the scope of an OKB can be consulted, analysed and possibly visualised.

---

[38] See Table 1 for an overview of the three models and how they are positioned on each dimension.

This model extends the previous models in the addition of services that allow user interaction with the data. Governance is *public*, insofar as the VRE is in control by public parties, but possibly developed by commercial parties who agree to open source software and public data licenses. The technical architecture is most likely (but not necessarily) *centralized*, which enhances performance of the VRE. An advantage of this model is that the VRE provides reference services that demonstrate the utility of the data stored in an OKB. For example, the VRE might include an Open Access Monitor or overviews of (numbers of) publications from participating institutes. A disadvantage is that costs are increased compared to the previous two models, since services need to be developed as well as sustained. Furthermore, an open question for this model is to what extent such a VRE could eventually replace existing commercial services for a number of use cases and thereby lead to cost savings.

Table 1. OKB models positioned on the different dimensions

Dimensions	**Open Knowledge API-standards**	**Open Knowledge Warehouse**	**Open Knowledge Research Environment**
Governance	Public-private (public standards, private APIs) Both top-down (standards) and bottom-up (API implementation)	Public (public control of warehouse) Top-down (central warehouse to which institutes deposit data)	Public (public control of environment) Top-down (central environment to which institutes deposit data)
Critical mass	Dependent on current platforms (critical mass exists with Pure)	At least five participating institutes + NWO	At least five participating institutes + NWO
Technical architecture	Federated	Centralized (but can be federated)	Centralized
Data scope	Identifiers, metadata, possibly full texts	Identifiers, metadata, possibly full texts	Identifiers, metadata, possibly full texts
Data quality	Quality assured by institutes	Quality assured and possibly enriched by central entity	Quality assured and possibly enriched by central entity
International connection	International standards	International standards, national warehouse	International standards, national environment
Service development	None	None (pure data layer)	Reference services or advanced VRE
Finances and funding	<€1 million (software development)	Start-up costs: €1-2 million (software development and hardware acquisition) Annual costs: €1 million (mainly personnel)	Start-up costs: €2-3 million (software development and hardware acquisition) Annual costs: €2 million (mainly personnel)

Dimensions	Open Knowledge API-standards	Open Knowledge Warehouse	Open Knowledge Research Environment
		Local institutional annual costs: <€0.5 million (personnel)	Local institutional annual costs: <€0.5 million (personnel)
Advantage	Can be achieved relatively quickly and cheaply, critical mass exists in institutes with Pure	Data quality can be harmonized and enriched	Data is immediately usable demonstrating utility of the data and an OKB
Risks	Ends up not truly "open" with limitations in APIs	More difficult to request additional data, thereby increased path-dependency. Ends up unused	More difficult to request additional data, thereby increased path-dependency. Services need strong usability focus

3.3 Stakeholder support for OKB models

Interviewees noted that **the API-standards model is insufficient to sustain an OKB**. *This model, however, provides the Warehouse model with data, and should consequently be pursued anyhow*. The main reason that this model is insufficient is that there is a risk that the API endpoints end up not being truly open. CRIS systems already offer API endpoints, but these pose limitations on the number of requests, the amount of data that can be pulled and what users can do with the data afterwards. Private parties thereby remain in control of the data. There is furthermore no scenario for data enrichments. This model in conclusion offers (too) little improvement over the current situation.

The majority of interviewees agreed that **the Warehouse model is sufficient and feasible to sustain an OKB**. The Warehouse model provides an open infrastructure to store open data that is critical to Dutch scholarship. The Warehouse model furthermore provides opportunities for further data enrichments, removing data redundancies, and analyses and assessments of national scholarship. Interviewees were, however, **critical whether the Warehouse model is sufficient to attract user engagement**. Without tools and services that demonstrate the utility of the data, there is a risk the data sits unused.

As a result, several interviewees noted that **the Research Environment model is desirable to sustain an OKB**. *This model should not be seen as a replacement but rather as an extension of the Warehouse model*. By offering a set of basic or perhaps even advanced tools and services, the Research Environment model demonstrates the utility of the data, attracts user engagement by addressing the use cases for the data, and allows further data enrichments for example through algorithmic classification of publications.

Appendix 1. Interviewees

#	User group	Name	Affliation
1.	Library IT	Alastair Dunning	TU Delft
2.	Library IT	Armand Guicherit	TU Delft
3.	Library IT	Enno Meijers	Koninklijke Bibliotheek
4.	Library IT	Erik Flikkenschild	Leiden University Medical Center
5.	Library IT	Henk Wals	DANS
6.	Library IT	Herbert van de Sompel	DANS
7.	Library IT	Lambert Heller	Technische Informationsbibliothek Hannover
8.	Library IT	Magchiel Bijsterbosch	SURF
9.	Library IT	Martijn Kleppe	Koninklijke Bibliotheek
10.	Library IT	Nick Veenstra	TU Eindhoven
11.	Library IT	Sören Auer	Technische Informationsbibliothek Hannover
12.	Institutional policy	Bianca Kramer	Utrecht University
13.	Institutional policy	Erna Sattler	Universiteit Leiden
14.	Institutional policy	Hans Ouwersloot	Maastricht University, DAIR
15.	Institutional policy	Jeroen Bosman	Utrecht University
16.	Institutional policy	Karin Maex	University of Amsterdam
17.	Institutional policy	Maurice Vanderfeesten	VU University Amsterdam
18.	National science policy	Darco Jansen	VSNU
19.	National science policy	Hans de Jonge	NWO
20.	National science policy	John Doove	SURF
21.	National science policy	Alexandra Vennekens	Rathenau instituut
22.	National science policy	Jeroen Heres	Rathenau instituut
23.	Researchers	Cameron Neylon	Curtin University, COKI
24.	Researchers	Egon Willighagen	Maastricht University
25.	Researchers	Ludo Waltman	Leiden University, CWTS
26.	Researchers	Natalia Manola	University of Athens, OpenAIRE
27.	Researchers	Paul Wouters	Leiden University
28.	Researchers	Rudi Bekkers	TU Eindhoven
29.	Researchers	Sarah de Rijcke	Leiden University, CWTS
30.	Researchers	Wilco Hazeleger	Utrecht University
31.	Private enterprise	Bo Alroe	Dimensions

#	User group	Name	Affliation
32.	Private enterprise	Hans Bos	Amazon Web Services

Contact:
Dialogic innovatie & interactie
Hooghiemstraplein 33-36
3514 AX Utrecht
Tel. +31 (0)30 215 05 80
www.dialogic.nl