Feasibility study
Open Knowledge Base

Commissioned by:
VSNU

Publication number:
2020.096-2108

Date:
Utrecht,
4 March 2021

Authors:
dr. Max Kemman
drs. Robbin te Velde

Table of contents

Executive summary

For Dutch see below.

The case for an Open Knowledge Base

The metadata on scholarly communications such as publications, datasets, software, educational material and communications aimed at the public, presents **critical information** on publicly funded scholarship that should be available without any restrictions. This metadata should therefore be easily findable, accessible and interoperable and reusable, where other users or service providers can create compelling use cases without barriers.

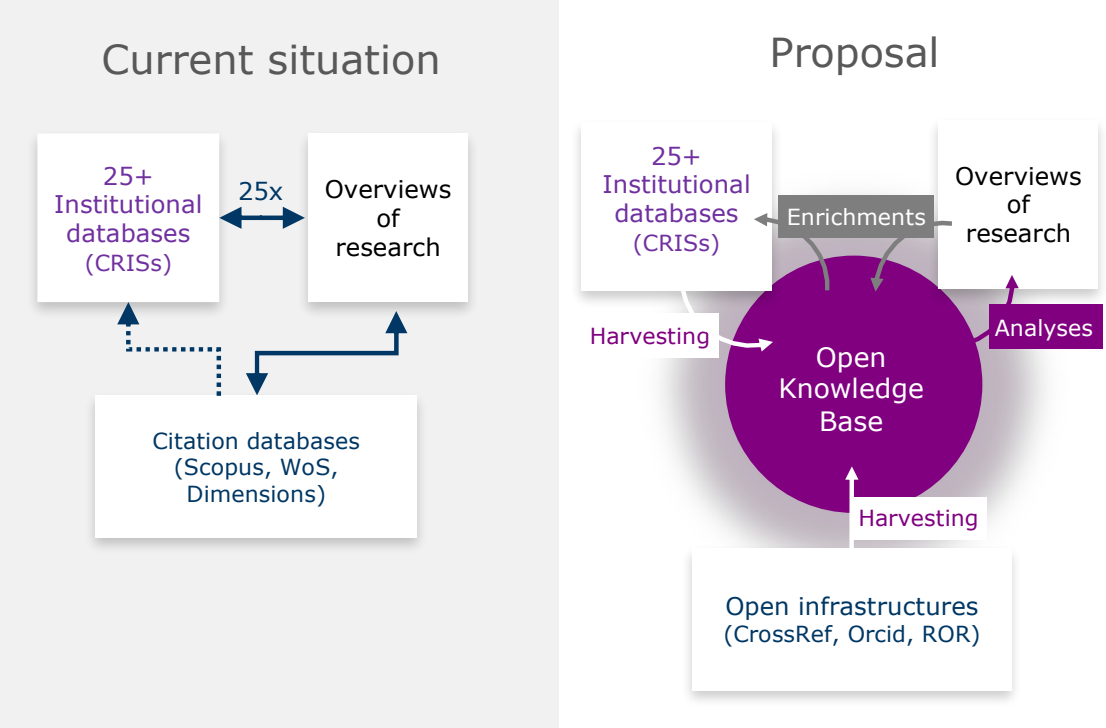


Figure representing the current situation and the OKB proposal. In the current situation overviews of research are generated using commercial citation databases and per institutional CRIS, of which over 25 are present in the Netherlands. In the OKB proposal metadata from the institutional CRISs is collected, enriched using metadata from open infrastructures and distributed back into CRISs to establish a feedback loop. Overviews of research can be generated on top of the OKB, from which additional enrichments may be fed back into the OKB.

However, the current landscape of research infrastructures presents two main issues. First, **academic independence is threatened** since overviews, evaluations and assessments of scholarship depend on citation databases governed by private enterprises. Second, where metadata on scholarly communications is available in public infrastructures, this **metadata is fragmented and lacking in quality and/or coverage**. As such, the core values of an open knowledge base (OKB) can be summarised as related to two concerns. First, to **protect academic independence** by opening up the metadata and metrics underlying assessments of scholarship and becoming less dependent on private enterprises for providing data and software. Second, to **improve and enhance the quality and coverage of metadata** available in the Dutch landscape of infrastructures on scholarly communications. By incorporating metadata on scholarly communications in an **open infrastructure** that not only digests but enriches and redistributes metadata, we posit that **an OKB may establish**

a feedback loop, see figure above. Through such a feedback loop, metadata coverage and quality are improved in the OKB by integrating, harmonising and enriching metadata from multiple sources including participating CRISs, open infrastructures (e.g., CrossRef, Orcid, ROR) and research intelligence services. These improvements and enrichments are fed back into institutional CRISs to improve metadata and subsequently overviews and reports at the local institutional level.

To this end, the OKB posits two proposals. First, a **technological proposal** of an open data layer that is interoperable and that prevents vertical integration by separating the data from the services. Second, a **governance proposal** to develop and maintain this technology and create buy-in from research institutes.

OKB models and scenarios for development

We identify three possible models for an OKB. These **models should not be seen as alternatives, but rather sequential models with increasing extensions of scope and complexity**. Later models depend on earlier models but offer more technological complexity and functionality.

- The **API-standards model** consists of a set of standards and guidelines of metadata that each institute or organisation should provide through an openly available API. This model is readily within reach, since Metis and Pure (the CRISs used by most research institutes) offer API-endpoints with the CERIF metadata standard. A risk of this model is, however, that is ends up not truly open in the sense of an API without restrictions to read, mix and share data, and that it is insufficient to establish a feedback loop.
- The **Warehouse model** consists of a centralised data warehouse where metadata is collected from the API-endpoints, deduplicated and harmonised. Metadata can furthermore be enriched and expanded from other open infrastructures (e.g., CrossRef, Orcid, ROR).
- The **Research Environment model** expands the Warehouse with the addition of research intelligence services and tools that **demonstrate the utility** of the data stored in the OKB and provides **references** for the development of alternative metrics. Some interviewees argued that such services are necessary to attract user engagement and institutional commitment. Furthermore, such services and tools may establish an **additional feedback loop** between the Warehouse and the services to expand or further improve metadata quality and coverage.

From these three models, we conclude that the Warehouse model is most feasible and desirable to facilitate a feedback loop. We subsequently identify four possible scenarios to develop an Open Knowledge Warehouse.

- 0-scenario: **maintain the current situation**. This scenario offers **open metadata in an open infrastructure** (i.e., NARCIS) but lacks a feedback loop or governance model to gain commitment to improve data quality and coverage. Research institutes subsequently remain dependent on commercial citation databases to gain overviews of scholarship. The 0-scenario thereby **does not sufficiently adequately address the core concerns** of academic independence or data coverage and quality.
- 1-scenario: **license a commercial product**. This scenario offers **open metadata in a closed infrastructure** that is feasible, usable and affordable and can be implemented relatively quickly. It is possible that metadata quality and coverage is improved through a feedback loop. Although the product may (and should) be licensed according to the Guiding Principles on Management of Research Information and Data to ensure the license meets public concerns, the infrastructure is dependent

on a commercial offering. As such, the 1-scenario **does not adequately address academic independence**.

- 2-scenario: **develop a bespoke solution**. This scenario builds an OKB by developing or reusing (parts of) open source software to offer **open metadata in an open infrastructure** that **establishes a feedback loop and governance to gain commitment**. Compared to the 1-scenario, a bespoke solution may (on the short-term) be **less usable and more costly**. It does however **fully address the core concerns** of academic independence and data quality and coverage.
- 3-scenario: **parallel maintaining, licensing and developing**. In this scenario **all three scenarios are pursued** for an agreed time period. In the final year an evaluation is conducted to **assess the impact of the scenarios** on the improvement of metadata quality and coverage. Furthermore, scenarios 1 and 2 can be compared to assess to what extent requests for metadata scope and development roadmap are satisfied or not and to assess technological complexity of bespoke development.

For the OKB as technological proposal we suggest a roadmap that identifies phases to design and develop all three models, with separate go/no go points with respect to the models as well as scenarios. Furthermore, for the OKB as governance proposal the roadmap identifies the need to create buy-in for three fundamental aspects; 1) long-term conditions, 2) feedback loop institutional CRISs-OKB, 3) feedback loop OKB-services. Long-term sustainability of an OKB requires four tasks:

- A **sustainable responsible team** with dedicated time and resources.
- **Sustaining of the technology** underlying the OKB.
- **Sustaining of buy-in** and creating buy-in beyond the initial critical mass, **organically positioning the OKB** in the national landscape, and **renewed discussions about definitions what counts** as scholarly output.
- Roadmap to develop necessary expertise and resources to sustain **potential competition** with private enterprises so that contract renewals are not from a position of dependence.

Recommendations

We make the following eight recommendations.

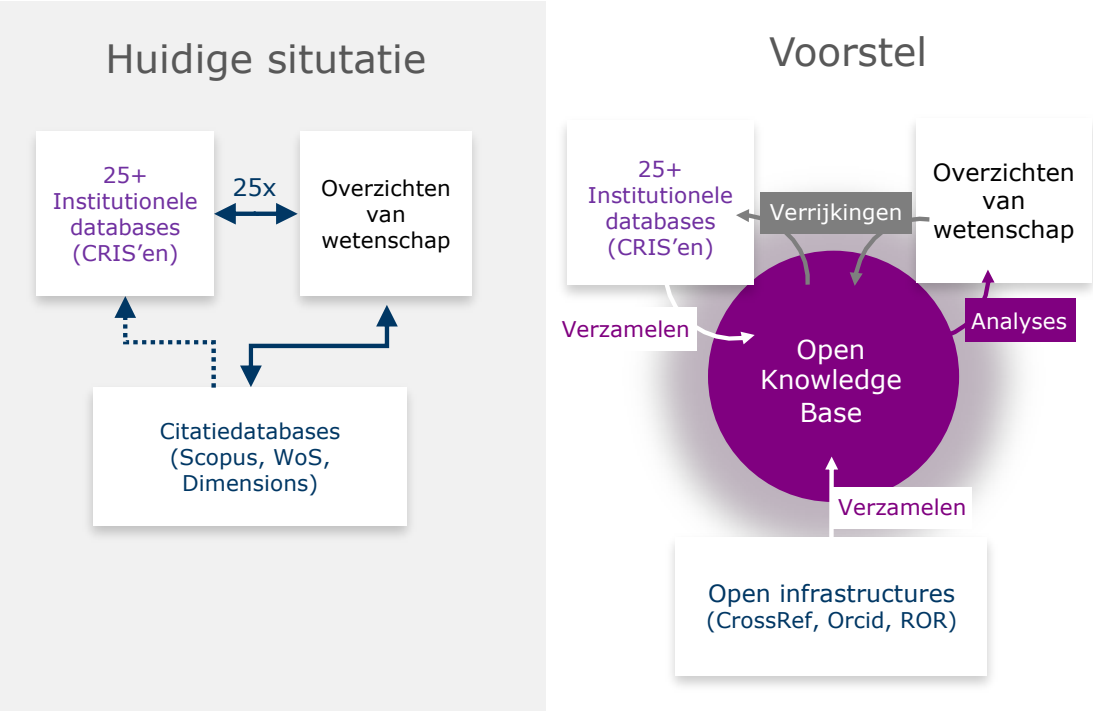
1. **Pursue (at minimum) the Warehouse model** to connect existing infrastructures in the Netherlands by collecting, storing, enriching and distributing metadata. As such an OKB addresses the concerns of academic independence and metadata quality and coverage.
2. **Collect in the Warehouse metadata on traditional objects** such as publications, grants, authors, institutes, funding agencies, **as well as non-traditional research output** such as datasets, software, scholarly communications aimed at the public and open educational resources. Include moreover **projects** to connect these objects in time.
3. **Establish a responsible team with a clear and strong mandate and dedicated time and resources** to make and pursue strategic decisions. This is possible by forming a working group within SURF, which subsequently provides legal basis for in-house development or tenders.
4. **Attract strong leadership** to lead the responsible team. To create buy-in, governance requires leadership that can **negotiate with the top level of research institutes** (rectorate and/or institutional policy). Prevent discussions and negotiations about an OKB to be limited to the library & IT level.

5. **Establish buy-in for metadata feedback loops** between institutional CRISs and the OKB (continuous metadata enrichment and enhancement) as well as between the OKB and research intelligence services (algorithmic enrichment).
6. **Aim at improving rather than replacing currently available systems** in the landscape of infrastructures on scholarly communications. This positioning should grow organically over time. With respect to NARCIS **it should be explored to what extent an OKB may (gracefully) supersede NARCIS as an infrastructure with sustainable funding and governance**.
7. **Initiate an OKB from the national level** but **position it in and ensure interoperability with the international landscape** of infrastructures on scholarly communications **by following international standards and data models**.
8. **Identify the necessary expertise and resources** to sustain an OKB and **create a roadmap to develop these conditions in the public sphere**. Even in case an OKB is developed by private enterprises or based on off-the-shelf products, it should remain possible to offer a **potentially competitive scenario** so that contract agreements and renewals are not from a position of dependence.

Managementsamenvatting

Waarom een open knowledge base?

De metadata over wetenschappelijke resultaten zoals publicaties, datasets, software, educatief materiaal en op het publiek gerichte communicatie, omvat **kritieke informatie** over met overheidsgeld gefinancierde wetenschap. Deze informatie zou zonder enige beperking beschikbaar zou moeten zijn. Deze metadata moet derhalve gemakkelijk vindbaar, toegankelijk, interoperabel en herbruikbaar zijn, zodat andere gebruikers of dienstverleners zonder belemmeringen overtuigende gebruiksmogelijkheden kunnen creëren.



Afbeelding met weergave van de huidige situatie en het OKB voorstel. In de huidige situatie worden overzichten van wetenschap verkregen uit commerciële citatie databases en per institutionele CRIS, waarvan er meer dan 25 aanwezig zijn in Nederland. In het OKB voorstel wordt metadata uit de institutionele CRIS'en verzameld, verrijkt met metadata uit open infrastructuren en terug gedistribueerd naar de CRIS'en om een feedback loop te bewerkstelligen. Overzichten van onderzoek kunnen worden verkregen op basis van de OKB, op basis waarvan additionele verrijkingen kunnen worden teruggevoerd in de OKB.

Er zijn echter twee belangrijke problemen in het huidige landschap van onderzoeksinfrastructuren. Ten eerste wordt de **academische onafhankelijkheid bedreigd** doordat overzichten, evaluaties en beoordelingen van wetenschapsbeoefening afhankelijk zijn van citatiedatabanken die door particuliere ondernemingen worden beheerd. Ten tweede, **metadata is gefragmenteerd en van onvoldoende kwaliteit en/of dekking** voor zover deze aanwezig is in openbare infrastructuren. De kernwaarden van een open knowledge base (OKB) kunnen dan ook worden samengevat tot twee overwegingen. Ten eerste, de **academische onafhankelijkheid beschermen** door de metadata en indicatoren die ten grondslag liggen aan de beoordeling van de wetenschap open te stellen en minder afhankelijk te worden van particuliere ondernemingen voor het verstrekken van gegevens en software. Ten tweede, **verbetering en uitbreiding van de kwaliteit en de**

dekking van de beschikbare metadata in het Nederlandse landschap van infrastructuren voor wetenschappelijke communicatie. Door metadata over wetenschappelijke resultaten op te nemen in een **open infrastructuur** die metadata niet alleen verzamelt maar ook verrijkt en verspreidt, stellen wij **dat een OKB een feedback loop tot stand kan brengen**, zie bovenstaande figuur. Door middel van een dergelijke feedback loop worden de metadata dekking en kwaliteit verbeterd in een OKB door het integreren, harmoniseren en verrijken van metadata uit meerdere bronnen waaronder deelnemende CRIS'en, open infrastructuren (bijv. CrossRef, Orcid, ROD) en research intelligence services. Deze verbeteringen en verrijkingen kunnen teruggevoerd worden aan institutionele CRIS'en zodat metadata en rapportages op lokaal institutioneel niveau worden verbeterd.

Zodoende presenteert de OKB twee voorstellen. Ten eerste een **technologisch voorstel** voor een open datalaag die interoperabel is en verticale integratie voorkomt door de data van de analyse diensten te scheiden. Ten tweede een **governance voorstel** om deze technologie te ontwikkelen en te onderhouden en draagvlak te creëren bij onderzoeksinstellingen.

OKB-modellen en scenario's voor ontwikkeling

We identificeren drie mogelijke modellen voor een OKB. Deze **modellen moeten niet worden gezien als alternatieven, maar als opvolgende modellen met toenemende scope en complexiteit**. Latere modellen zijn hierbij afhankelijk van eerdere modellen, maar bieden meer technologische complexiteit en functionaliteit.

- Het **API-standaarden model** bestaat uit een set standaarden en richtlijnen van metadata die elk instituut of organisatie zou moeten aanbieden via een openbare API. Dit model ligt reeds binnen handbereik, aangezien Metis en Pure (de CRIS'en die door de meeste onderzoeksinstellingen worden gebruikt) API-eindpunten bieden met de CERIF-metadatastandaard. Een risico van dit model is echter dat het uiteindelijk niet echt open is in de zin van een API zonder beperkingen om gegevens te lezen, aan te passen en te delen, en dat het onvoldoende is om een feedbackloop op te zetten.
- Het **Warehouse model** bestaat uit een gecentraliseerd data warehouse waarin metadata van de API's worden verzameld, gededupliceerd en geharmoniseerd. Metadata kunnen bovendien worden verrijkt en uitgebreid vanuit andere open infrastructuren (bijv. CrossRef, Orcid, ROR).
- Het **Onderzoeksomgeving model** breidt het Warehouse uit met de toevoeging van diensten en research intelligence services die het nut aantonen van de gegevens die in de OKB zijn opgeslagen en referenties bieden voor de ontwikkeling van alternatieve indicatoren. Sommige gesprekspartners voerden aan dat dergelijke diensten noodzakelijk zijn om interesse van gebruikers alsook institutionele betrokkenheid te vergroten. Bovendien kunnen dergelijke diensten en software een extra feedbackloop tussen het Warehouse en de diensten tot stand brengen om de kwaliteit en de dekking van de metadata verder te verbeteren.

Uit deze drie modellen concluderen we dat het Warehouse-model het meest haalbaar en wenselijk is om een feedback loop te faciliteren. Vervolgens stellen wij vier mogelijke scenario's vast voor de ontwikkeling van een Open Knowledge Warehouse.

- 0-scenario: **in stand houden van de huidige situatie**. Dit scenario biedt **open metadata in een open infrastructuur** (d.w.z. NARCIS), maar mist een feedback loop of een governance model om commitment te krijgen voor het verbeteren van kwaliteit en dekking van de metadata. Onderzoeksinstituten blijven daardoor afhankelijk van commerciële citatiedatabases om overzichten van de wetenschap te

krijgen. Het 0-scenario **komt onvoldoende tegemoet aan de kernproblemen** van academische onafhankelijkheid en dekking en kwaliteit van de gegevens.

- 1-scenario: **afnemen van een commercieel product**. Dit scenario biedt **open metadata in een gesloten infrastructuur** die haalbaar, bruikbaar en betaalbaar is en relatief snel kan worden geïmplementeerd. Hierbij kan het mogelijk zijn om de kwaliteit en dekking van de metadata te verbeteren middels een feedback loop. Hoewel het product kan (en moet) worden afgenomen volgens de Guiding Principles on Management of Research Information and Data om publieke waarden te borgen in de licentie, wordt de infrastructuur afhankelijk van een commerciële aanbieder. Het 1-scenario **beantwoordt daarmee niet aan academische onafhankelijkheid**.
- 2-scenario: **ontwikkel een oplossing op maat**. Dit scenario bouwt een OKB door het ontwikkelen of hergebruiken van (delen van) open source software om **open metadata aan te bieden in een open infrastructuur** die een **feedback loop en governance opzet om commitment te verkrijgen**. In vergelijking met het 1-scenario kan een op maat gemaakte oplossing (op korte termijn) minder bruikbaar en duurder zijn. Het **komt echter volledig tegemoet aan de kernproblemen** van academische onafhankelijkheid en kwaliteit en dekking van de gegevens.
- 3-scenario: **parallel in stand houden, afnemen en ontwikkelen**. In dit scenario worden **alle drie de scenario's gedurende een afgesproken periode gevolgd**. In het laatste jaar wordt een evaluatie uitgevoerd voor de **beoordeling van de impact van de scenario's** op de verbetering van de kwaliteit en de dekking van de metadata. Daarbij kunnen scenario's 1 en 2 worden vergeleken om na te gaan in welke mate al dan niet wordt voldaan aan verzoeken inzake de scope van metadata en functionele ontwikkeling, en om de technologische complexiteit van de ontwikkeling op maat te beoordelen.

Voor de OKB als technologisch voorstel stellen wij een stappenplan voor dat fasen aangeeft voor het ontwerp en de ontwikkeling van alle drie de modellen, met afzonderlijke go/no go-punten met betrekking tot de modellen en de scenario's. Bovendien identificeert de routekaart voor het OKB als governance voorstel de noodzaak om draagvlak te creëren voor drie fundamentele aspecten; 1) lange termijn voorwaarden, 2) feedback loop tussen institutionele CRIS'en en de OKB, 3) feedback loop tussen de OKB en diensten. De duurzaamheid van een OKB op de lange termijn vereist vier taken:

- Een **duurzaam verantwoordelijk team** met toegewijde tijd en middelen.
- **Duurzaam onderhoud van de technologie** die aan de OKB ten grondslag ligt.
- Het **onderhouden van draagvlak** en het creëren van draagvlak na de initiële kritieke massa, het **organisch positioneren van de** OKB in het nationale landschap, en **hernieuwde discussies over definities wat telt** als wetenschappelijke output.
- Routekaart om de benodigde expertise en middelen te ontwikkelen om **potentiële concurrentie** met particuliere ondernemingen in stand te houden, zodat contractverlengingen niet vanuit een positie van afhankelijkheid plaatsvinden.

Aanbevelingen

Wij doen de volgende acht aanbevelingen.

1. **Streef (ten minste) het Warehouse model na** zodat bestaande infrastructuren in Nederland met elkaar worden verbonden door metadata te verzamelen, op te slaan, te verrijken en te distribueren. Zo komt een OKB tegemoet aan de zorgen over academische onafhankelijkheid en de kwaliteit en dekking van metadata.

2. **Verzamel in het Warehouse zowel metadata over traditionele objecten** zoals publicaties, beurzen, auteurs, instituten, financieringsinstanties, **als metadata over niet-traditionele onderzoeksoutput** zoals datasets, software, op het publiek gerichte wetenschappelijke communicatie en open onderwijsmiddelen op. Neem daarnaast **projecten** op om deze objecten in de tijd met elkaar te verbinden.
3. **Stel een verantwoordelijk team samen met een duidelijk en sterk mandaat en toegewijde tijd en middelen** om strategische beslissingen te nemen en uit te voeren. Dit kan door het vormen van een werkgroep binnen SURF, die een juridische basis biedt voor eigen ontwikkeling of aanbestedingen.
4. **Trek sterk leiderschap aan** om het verantwoordelijke team te leiden. Om draagvlak te creëren is voor de governance leiderschap nodig dat kan **onderhandelen met de top van de onderzoeksinstellingen** (rectoraat en/of instellingsbeleid). Voorkom dat discussies en onderhandelingen over een OKB beperkt blijven tot het bibliotheek- & IT-niveau.
5. **Creëer draagvlak tot stand worden voor metadata feedback loops** tussen institutionele CRIS'en en de OKB (continue metadata verrijking en verbetering) alsmede tussen de OKB en research intelligence services (algoritmische verrijking).
6. **Richt doelstellingen op het verbeteren van bestaande systemen** in het landschap van infrastructuren voor wetenschappelijke communicatie **en niet op vervangen**. De positionering moet in de loop van de tijd organisch groeien. Met betrekking tot NARCIS **moet worden verkend in hoeverre een OKB (geleidelijk) NARCIS kan vervangen als een infrastructuur met duurzame financiering en bestuur**.
7. **Initieer een OKB vanuit het nationale niveau**, maar **positioneer deze in en zorg voor interoperabiliteit met het internationale landschap** van infrastructuren voor wetenschappelijke communicatie **door internationale standaarden en datamodellen te volgen**.
8. **Identificeer de benodigde expertise en middelen** om een OKB in stand te houden en **ontwerp een routekaart om deze voorwaarden in de publieke sfeer te ontwikkelen**. Zelfs wanneer een OKB wordt ontwikkeld door particuliere ondernemingen of gebaseerd is op bestaande producten, moet het mogelijk blijven een **potentieel concurrerend scenario** te bieden, zodat contractovereenkomsten en verlengingen niet vanuit een positie van afhankelijkheid plaatsvinden.

1 Introduction

1.1 Problem statement

Open Science is propelling deep-seated change in the way scientific endeavour is conducted, recognised and rewarded. Not only does Open Science aim to establish more open and transparent research methods but this transition also concerns increased cooperation and transparency of research, publishing and research assessments.[1] Embedding Open Science demands appropriate infrastructures. Such infra- structures 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.

Science and higher education are further becoming increasingly digitalised. In this digitalisation and increasing dependence on digital platforms, higher education risks becoming increasingly dependent on large technology companies. In a recent opinion article, the rectors of the Dutch universities called for more critical reflection on this **threat of dependency** (and thereby threat to academic independency) and for a recognition of the shared responsibility of public institutes to ensure digital platforms conform to public values and norms.[2] Just as large technology companies, large publishing companies are increasingly moving to service-oriented models based on data ('Platform economy')[3]. 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 which may offer an "exit-strategy" for universities in case they do not extend current contracts with large publishing companies. The rectors of Dutch universities as well as the Chief Innovation Officer of SURF therefore suggest that cooperation with large private enterprises should be paralleled by independent developments towards open alternatives that may offer **competitive potential**.[4]

As such, early 2020 the VSNU (the Association of Universities in the Netherlands) established the Dutch taskforce on Responsible Management and Research Information and Data[5] (from here on, the taskforce) to develop guidelines on how Dutch research institutions can

---

[1] What is Open Science? [accelerateopenscience.nl]

[2] Maex et al. (2020). Digitalisering bedreigt onze universiteit. Het is tijd om een grens te trekken. Opinion article in *De Volkskrant* [volkskrant.nl]; see also KNAW (2021). *Academische vrijheid in Nederland – een begripsanalyse en richtsnoer*. Amsterdam, KNAW, pp. 40-42.

[3] 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; Björn Brembs et al. (2020). Auf einmal Laborratte. *Frankfurter Allgemeine.* [zeitung.faz.net]

[4] SURF loopt niet alleen aan de hand van grote tech-spelers (2021). *ScienceGuide* [scienceguide.nl]; see also van Dijck, J., Poell, T., & de Waal, M. (2016). *De platformsamenleving. Strijd om publieke waarden in een online wereld*. Amsterdam University Press. [doi:10.5117/9789462984615], chapter 6 for a discussion of government and public institutes in roles as users, regulators or developers or digital platforms.

[5] VSNU (2020). Dutch Taskforce on Responsible Management of Research Information and Data. [vsnu.nl]

collaborate with commercial entities, which where (partially) implemented in the contract with Elsevier[6]. These proposed guidelines[7] (which are under revision following a public consultation) encompass the following principles:

1. The ownership of research output and related metadata resides with the institutions and researchers.
2. Services that collect, interact with or use scholarly output in any shape or form must facilitate complete, non-discriminatory and enduring access to primary metadata and derivative data without functional, technical, legal or financial limitations.
3. All metadata held must have transparent, trusted provenance.
4. Interoperability should be ensured by standardised scholarly metadata which is accessible and separated from associated services and tools.
5. Services can be outsourced to different market players as long as monopolies are prevented.
6. Governance should foster an inclusive sustainable decision-making process.

Another concern with the current situation is that 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 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. To this end, the taskforces proposes that an open knowledge base (OKB) could address such concerns and improve compliance with the above guidelines[8].

1.2 The Open Knowledge Base proposal

An OKB is an infrastructure through which metadata of scholarly communications is made accessible.[9] This metadata is collected from metadata providers such as institutional CRIS systems to provide a single point of access to the metadata within scope (in this case, the scholarly output of the Dutch academic community, notably but not exclusively research universities, academic hospitals, and academy research institutes[10]) and enable the development of subsequent services and tools.

1.2.1 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 metadata on scholarly communications, such as publications, are as such arguably **critical information** on scholarship that should be available to the public without any restrictions. This metadata should therefore be easily findable, accessible and interoperable and reusable[11], where other users or service providers

---

[6] VSNU (2020). Dutch research institutions and Elsevier initiate world's first national Open Science partnership. [vsnu.nl]

[7] VSNU (2020). Guiding Principles on Management of Research Information and Data. [vsnu.nl]

[8] VSNU (2020). Towards an Open Knowledge Base. *Guiding Principles on Management of Research Information and Data.*

[9] Dunning, Vanderfeesten, De Rijcke, Bijsterbosch, Jansen (2020). What is an Open Knowledge Base anyway? [openworking.wordpress.com]

[10] The Academy institutes [knaw.nl]

[11] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 [doi:10.1038/sdata.2016.18]

can create compelling use cases without barriers. This may furthermore make datasets, software, lab notebooks or other output related to open science more easily discoverable through links with publications, researchers and research programmes.

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)[12] reports. A limitation of the **SEP and KUOZ reports** with the current situation is that institutes report aggregated data, while the data remains in separate 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). By providing access to the underlying data in a transparent way, differences in such interpretations can be analysed, identified and possibly resolved through mutual negotiation. 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[13] 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.

In the public sector, discussions related to open science and open access have gained momentum in recent years.[14] For instance, the current way in which academic output is measured is increasingly criticised, such as debates on the recognition of academic impact 'Erkennen & Waarderen'.[15] 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.[16] An OKB could facilitate future **evaluations of the different methods of funding**.

A second core argument for an OKB is that where metadata on scholarly communications is available in public infrastructures, this **metadata is fragmented and lacking in quality and/or coverage**. 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.[17] The Dutch government subsequently regularly sets agendas prioritising specific research fields at the expense of others, for examples in

---

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

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

[14] Maex et al. (2020). Digitalisering bedreigt onze universiteit. Het is tijd om een grens te trekken. Opinion article in *De Volkskrant* [volkskrant.nl].

[15] VSNU, NFU, KNAW, NWO, ZonMw (2019). Ruimte voor ieders talent: naar een nieuwe balans in het erkennen en waarderen van wetenschappers. Position paper.

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

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

programmes such as the Dutch Research Agenda[18] or Topsectoren[19]. Approximately two-thirds (€655 million) of the annual budget of the national science funder NWO is allocated for thematic research.[20] 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).[21] Furthermore, these services predominantly index journal articles and do not adequately cover other scholarly output such as books, code, data, et cetera.[22] Finally, these services no longer truly fit the Dutch commitment to alternative assessments of Dutch scholarship.[23] Yet public metadata repositories such as NARCIS have been found lacking to develop overviews or analyses of Dutch scholarship (see §2.1.5 below).

In conclusion, the core values of an OKB can be summarised as two concerns. First, to **protect academic independence** by opening up the metadata and metrics underlying assessments of scholarship. Second, to **improve and enhance the quality and coverage of metadata** available in the Dutch landscape of infrastructures on scholarly communications.

1.2.2 OKB as a technological proposal

Figure 1 provides a sketch of what an OKB may entail on a technical level, describing three separate layers. First and foremost, a **primary data layer** that aggregates data from institutional CRISs and other data sources such as CrossRef, ORCID, or other infrastructures containing data on Dutch scholarly communications. Second, an OKB includes a **secondary data layer** in which the data from the primary layer is filtered and curated to fit identified use cases. An important aspect for this secondary layer is **deduplication** of records where metadata on for example an article originating from more than one institutional CRIS or other scholarly infrastructure is identified to refer to the same scholarly output. Third and finally, an OKB may (but need not, this is an aspect of our exploration of the scope of an OKB in this report) include a **tools layer** which provides tools and services to analyse and use the data in the OKB.

---

[18] [nwo.nl]

[19] [topsectoren.nl]

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

[21] 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.

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

[23] 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.

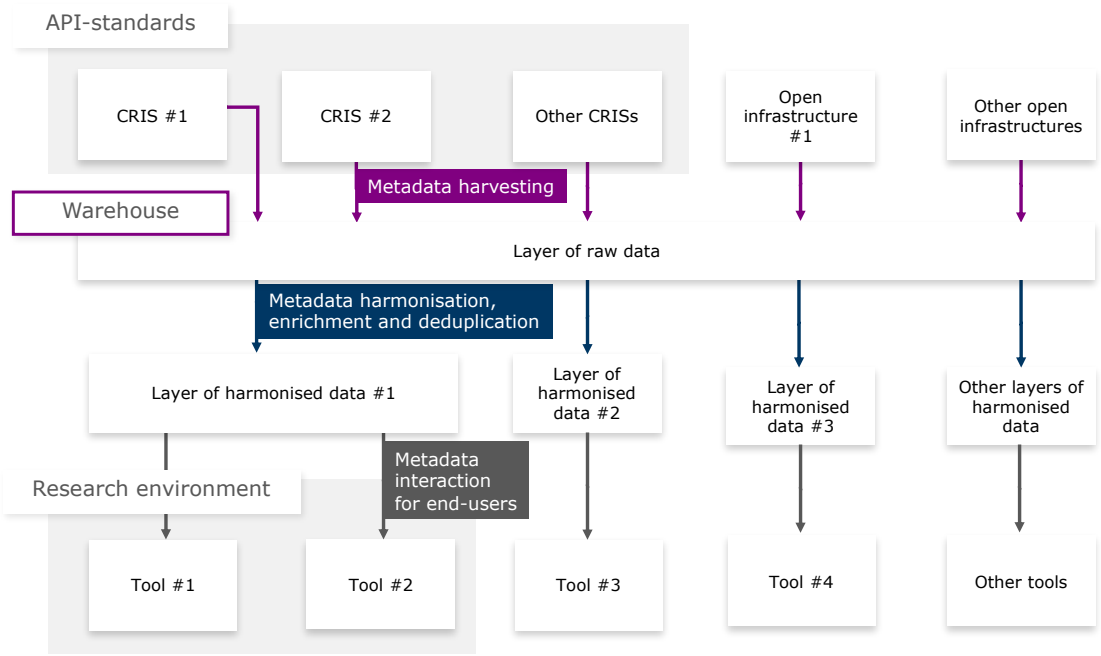


Figure 1. Sketch of what an open knowledge base might entail as a technological proposal in three levels: API-standards, Warehouse and Research Environment (elaborated in section 3.2). Metadata is collected from institutional CRISs as well as open infrastructures. Multiple layers of harmonised data can be created from this collected data to facilitate specific use cases; these layers may be developed as part of an OKB or may be developed by other parties. Finally, tools are developed that enable interaction with the metadata by end-users; these tools may be developed as part of an OKB or may be developed by other parties. Sketch inspired by earlier sketches developed at the OKB Hackathon (see section ***Error! Reference source not found.****).*

Two aspects of an OKB are deemed central. Firstly, an OKB facilitates that research institutes start using a **single data infrastructure**. This is very challenging, but there are also advantages from building on both the shared human knowledge as well as the technical resources at research institutes. 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.

The purposes of aggregation of scholarly metadata and deduplication into a single catalogue of Dutch scholarly communications is comparable to the National Academic Research and Collaborations Information System (NARCIS)[24] maintained by DANS, which has existed since 2004. However, NARCIS currently faces several limitations with respect to data quality and coverage that an OKB may overcome (see §2.1.5 below).

1.2.3 OKB as a governance proposal

Since an OKB is centrally concerned with public ownership of metadata on scholarly communications to secure academic independency and transparency of scholarship as well as the quality of analyses of that scholarship, an OKB furthermore includes a **governance**

[24] About NARCIS [narcis.nl]

proposal. The introduction of an infrastructural component that requires development, maintenance and commitment from research institutes entails that an OKB needs governance and needs clarification of ownership. How to design governance and development in the short-term as well as governance and maintenance in the long-term are still open questions. The elaboration of this aspect of governance is another aspect where the proposal for an OKB may extend or supersede NARCIS.

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.

1.3 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 considered 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?

1.4 Process of the feasibility study

The VSNU and the taskforce commissioned[25] Dialogic in August 2020 to assess the feasibility of an open knowledge base and investigate the research questions outlined above. In the

[25] VSNU (2020). Terms of reference – Feasibility study on an open knowledge base for the Dutch research community. [vsnu.nl]

period of September to December 2020 we conducted interviews with 44 stakeholders. We classified twenty interviewees as Library & IT, five as institutional policy, three as national science policy, eleven as researchers, and five 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[26] and Curtin Open Knowledge Initiative[27] in November 2020.

During the feasibility study progress reports were delivered to the supervisory committee, consisting of members of the taskforce. In December 2020, a progress report (v0.77)[28] was shared in a **public consultation** with user groups to receive input on how they perceived the feasibility of an OKB and the identified use cases. We received twelve replies, mainly from Dutch Library & IT. Respondents argued that the potential value of an OKB should not be sought in use cases where user groups directly engage with an OKB. Instead, respondents noted that an OKB should be more clearly positioned in the current landscape of Dutch infrastructures on scholarly communications to identify how an OKB may improve this landscape. As a result, chapter 2 was significantly rewritten to better identify what use cases an OKB should facilitate.

1.5 Purpose and structure of this report

This report discusses our findings pertaining to the above key goal of the feasibility study and addresses the research questions listed above. The starting point of this report is that an Open Knowledge Base (OKB) is desired by Dutch research institutes[29] and explores what an OKB could entail and what factors impact the feasibility and roadmap of development and implementation of an OKB. The purpose of this report is notably not to explore the desirability of an OKB or alternative solutions, but to explore the points of decision towards an OKB if desired.

In Chapter 2 we map the current Dutch landscape of infrastructures on scholarly communications and describe how an OKB may be positioned within this landscape. Furthermore, we explore some of the limitations of the current situation and how these may be resolved by an OKB. In Chapter 3 we analyse the possible characteristics of an OKB and how these lead to *three* different models for an OKB. We furthermore explore four different scenarios to pursue one of these models. In Chapter 4 we will discuss the roadmap for development and implementation of each possible model. Finally, in Chapter 5 we present our conclusions and recommendations to ensure the feasibility of an OKB.

[26] [cwts.nl]

[27] [openknowledge.community]

[28] Feasibility study Open Knowledge Base - version for consultation [doi:10.5281/zenodo.4304334]

[29] See VSNU (2020). Towards an Open Knowledge Base. *Guiding Principles on Management of Research Information and Data*.

2 Positioning an OKB in the landscape of existing infrastructures

In this chapter we consider how an OKB may be positioned in the landscape of infrastructures on scholarly communications and how this landscape is experienced by four user groups: researchers, library & IT, national science policy, institutional policy. In section 2.1 we map the current landscape and identify problems that user groups currently experience. In section 2.2 we describe how an OKB may be positioned within the landscape and suggest how this positioning may resolve some of the identified problems.

2.1 The current landscape of infrastructures on scholarly communications

To understand the demands of the different user groups with respect to an open knowledge base, the first concern is what problems exist in the current Dutch landscape of infrastructures on scholarly communications. Based on the interviews and desk research, we have mapped the current landscape to identify gaps and problems where an OKB may provide a solution. This map is shown is Figure 2 and shows how the metadata is generated by the user groups researchers and library & IT on the left side and moves through different systems towards institutional and national policy makers who make use of this metadata. Note that this map is not comprehensive and for readability may leave out some connections. Important to note is that while this map includes a single box for institutional CRIS, in reality this represents 24 research institutes that use Pure from Elsevier[30], four research institutes that use Metis developed and maintained by Radboud University Nijmegen, and one university that uses Converis from Clarivate[31] (i.e., Leiden University).

[30] Pure Clients [elsevier.com]

[31] Converis [clarivate.com]

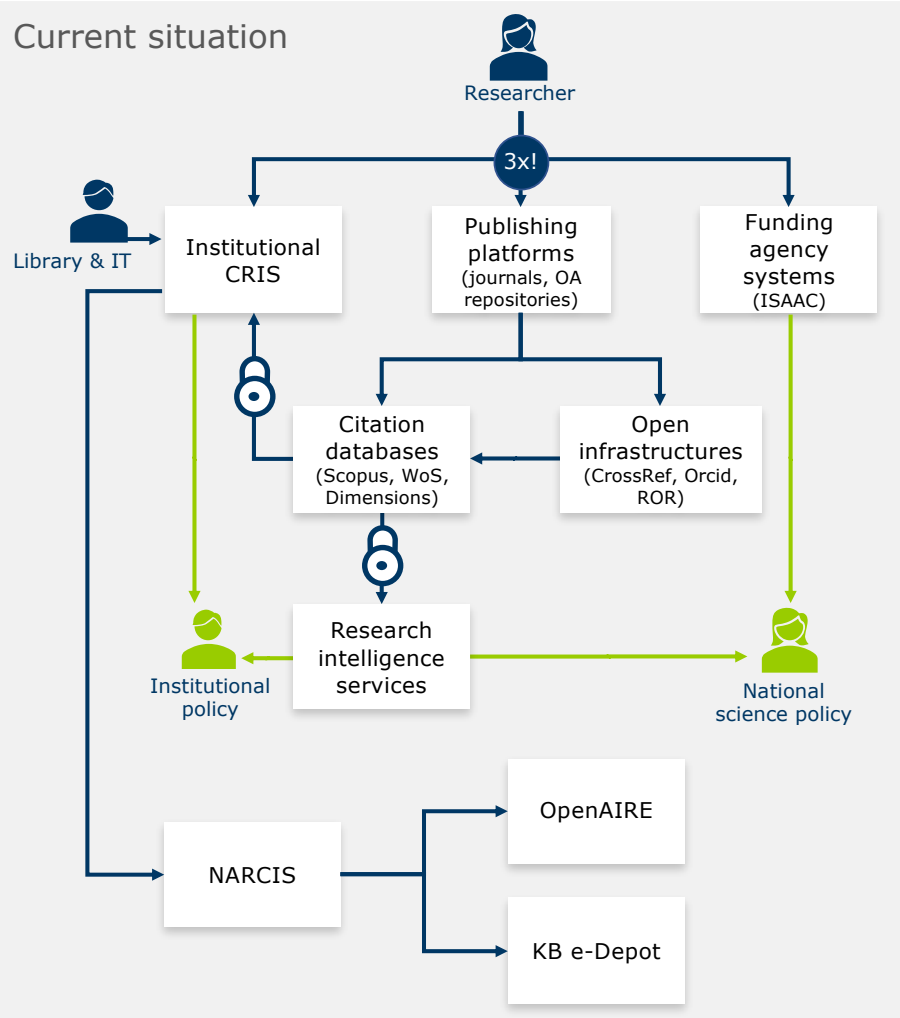


Figure 2. Map of current Dutch landscape of infrastructures on scholarly communications. Blue lines indicate the metadata pipeline moving from top to bottom. Green lines indicate where metadata is used for overviews of scholarship. Locks indicate pipelines that are dependent on licenses from private enterprises. This map shows three problems with the current situation. First, researchers have to enter data three times in an institutional CRIS, a publishing platform and the administrative system of a funding agency. Second, overviews on research are strongly dependent on commercial citation databases. Third, the metadata pipeline to open infrastructures such as NARCIS and OpenAIRE is a one-way route with no feedback on data quality and coverage.

2.1.1 Researchers

For the researcher user group, a problem with the current landscape that was described repeatedly in the interviews and that becomes apparent in Figure 2 is that **researchers have to provide metadata on publications several times**. When researchers publish in a journal or deposit data at an open access repository (e.g., DANS EASY[32], Zenodo[33], figshare[34]), they first provide the appropriate metadata (title, author(s), affiliations, keywords, abstract) to the administrative system of the publishing platform. Second, researchers provide this same metadata to their institutional CRIS (without affiliation, which

[32] [easy.dans.knaw.nl]

[33] [zenodo.org]

[34] [figshare.com]

is covered by the CRIS, but with journal title, volume, issue and possibly DOI). This way, the institution gains an overview of the scholarly output of its researchers. Third, researcher provide this metadata to the administrative system of the funder (in the case of NWO this system is ISAAC) so the funder gains an overview of the scholarly output enabled by its grants. While clear incentives exist for researchers to provide this metadata, several interviewees noted that due to the necessity to manually enter the same data several times, data quality is lacking. Data is entered close to the deadline rather than in real-time, leading to gaps, and errors may be entered. Furthermore, in the case of collaborations of researchers at multiple institutes, the same metadata has to be entered by researchers in their respective institutional CRIS. **For researchers, the main potential benefit of an OKB is to decrease the amount of administrative work**. This is to some extent alleviated by the Pure CRIS which can notify researchers of publications linked to them in the citation database Scopus, which ingests the data from the journals. A good example of alleviating this problem with respect to funding agency systems is the Finnish VIRTA system[35], which offers integration with the national funder's application system. During a grant application, researchers can select publications from VIRTA to add these as part of the proposal. Since VIRTA retrieves these publications from the institutional CRIS systems, researchers are incentivised to keep their publications up to date for grant proposals.

Furthermore, an additional benefit for researchers may be that an OKB could facilitate better registration of research output beyond journal publications, including metadata on research data, software, open educational resources, etc. In this sense an OKB may enhance information systems to support metrics related to 'Erkennen & Waarderen'[36].

2.1.2 Library & IT

For the library & IT user group, the main problem with the current landscape mentioned in the interviews is that each institute maintains their own institutional CRIS content. When a librarian at one institute enters or enriches metadata in their CRIS, this information is not shared with other CRISs. **CRISs across institutions consequently 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. **For library & IT, the main potential benefit of an OKB is to share metadata between institutes by feeding data from an OKB back into institutional CRISs**.

2.1.3 National science policy

For the national science policy user group, an important concern mentioned in the interviews is that the metadata entered in ISAAC is regularly insufficiently comprehensive or lacking in quality. The cause of this problem is identified above for the researcher user group, who have to maintain data in multiple administrative systems. 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. **An OKB may aid harmonising of the record of scholarly communications by making such differences in interpretation of what counts as a publication explicit and visible for discussion**. Such differences in interpretations are currently hidden, since

[35] VIRTA in English [wiki.eduuni.fi]

[36] VSNU et al. (2019). Ruimte voor ieders talent.

institutes report figures in aggregate without disclosing the underlying data. Furthermore, insofar as national science policy makers depend on research intelligence services based on commercial citation databases, this renders the overview of scholarship dependent on the interpretation of large publishers of what counts as scholarly communications. A consequence of these problems is that **funders currently have a limited overview of the scholarship that is produced** as a result of funding grants. Another concern for the national science policy user group is that in gaining an overview of scholarship and assessing quality, for many policy questions this is currently dependent on research intelligence services integrated with citation databases from private enterprises (represented by the closed access icons in Figure 2). This is especially of concern in the context of debates surrounding academic independence, where strategic decision making is currently dependent on metrics developed by private enterprises.

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').[37] 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 beyond articles and citations.[38] For example, science policy makers might wish to monitor how Dutch scholarship contributes to the United Nation's sustainable development goals (SDG).[39] For the Netherlands, metrics of national interest may monitor how Dutch scholarship contributes to mission-oriented research and key enabling technologies.[40]

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".[41] 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.[42] Finally, recent discussions 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.[43]

---

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

[38] This aspect arguably relates as well to the Erkennen & Waarderen debate.

[39] 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.

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

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

[42] 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].

[43] 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.

An OKB could provide the infrastructure for more reliable and relevant (next generation) metrics that better align with national priorities and strategies, 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.[44]

2.1.4 Institutional policy

Interviewees from institutional policy however 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.

The **main opportunity of an OKB for institutional policy is then to improve the quality and coverage of metadata**. By improving the quality and coverage of institutional CRIS systems, an OKB leads to further **improvements and enhancements of the SEP and KUOZ reports** that institutes are obligated to produce. Furthermore, since an OKB contains the same metadata for other institutes, it becomes feasible to **position one's own institute** in the landscape of Dutch research institutes. In this context, 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 (traditional) scientific impact as well as societal impact such as SDGs or mission-oriented research and key enabling technologies.[45] Furthermore, institutional managers could develop their own metrics and assess how their institute compares to other institutes depending on what they find strategically important.[46]

2.1.5 Metadata preservation

Finally, while not defined as a user group, is it important to consider how existing infrastructures on scholarly communications act to preserve and distribute metadata on Dutch scholarly communications. NARCIS harvests the institutional CRISs to create an overview of Dutch scholarship. NARCIS underlies the National Library's e-Depot which preserves all literature produced by Dutch people, where publications in NARCIS from Dutch scholars or scholars with a sustainable Dutch position are preserved. Furthermore, NARCIS metadata is shared with OpenAIRE[47], the European open science infrastructure for scholarly communications.

---

[44] 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]).

[45] Such metrics and analyses may be particularly relevant for Universities of Applied Sciences where research is more strongly aimed at societal impact, see Sarah K. Coombs & Ingeborg Meijer (2021). Towards Evaluating the Research Impact made by Universities of Applied Sciences, *Science and Public Policy*, [doi:10.1093/scipol/scab009]; Vereniging Hogescholen (2016). *Onderzoek met impact – Strategische onderzoeksagenda hbo 2016-2020.*

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

[47] [openaire.eu]

Interviewees noted that the quality and coverage of metadata in NARCIS and subsequently OpenAIRE is found lacking. While on a technological level the harvesting is successfully implemented, the **problems earlier in the metadata chain identified at the point of data input by researchers and library & IT are reinforced**. Furthermore, since the user groups do not directly make use of NARCIS or subsequent infrastructures, interviewees noted there is **little commitment to improve or sustain these infrastructures**.

2.2 Positioning of an OKB in the landscape

Figure 3 shows how an OKB may be positioned in the landscape of infrastructures on scholarly communications. In the following paragraphs, we describe how for each user group an OKB might alleviate or solve the problems identified in the previous section. It should be noted that we have positioned the **OKB as an infrastructure that is largely invisible to the user groups**, but that underlies and improves the systems that are readily available and in use.

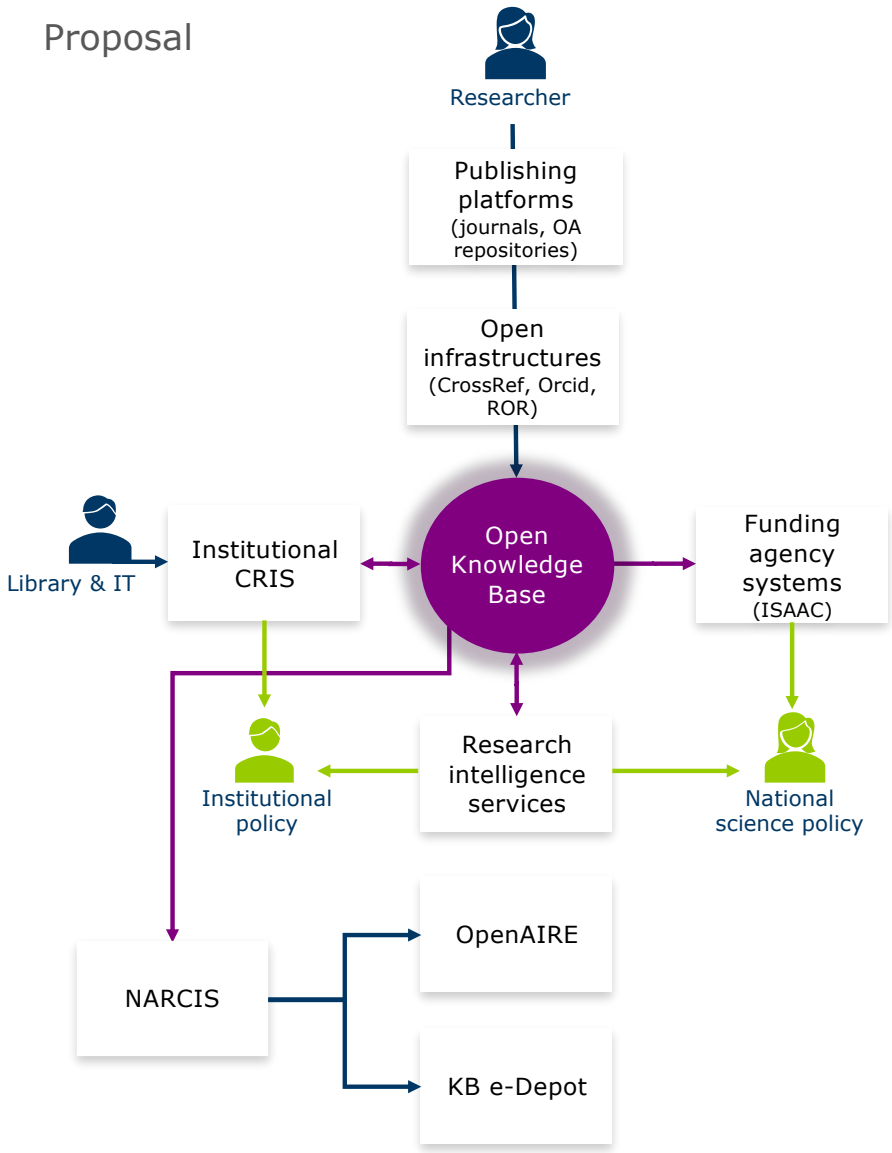


Figure 3. Map of Dutch landscape positioning an Open Knowledge Base. Blue lines indicate the metadata pipeline moving from top to bottom. Green lines indicate where metadata is used for overviews of scholarship. Purple lines indicate how an OKB may ingest and distribute metadata from and to other systems. Note especially the feedback loop between institutional CRISs, the OKB and research intelligence services in which metadata is moved to increasingly enhance quality and coverage. Note also that user groups in principle do not directly interact with the OKB. Metadata pipelines present in Figure 2 may still exist but are hidden to enhance clarity of the map. Compared with Figure 2, this map demonstrates three significant improvements. First, researchers only have to enter metadata once. Second, overviews of research are not dependent on commercial citation databases. Third, the metadata pipeline to other open infrastructures ingests metadata improvements achieved in the feedback loop established by the OKB.

2.2.1 Researchers

For researchers we identified the main potential benefit of an OKB as to facilitate and decrease the amount of administrative work. As shown in Figure 3, **an OKB supports the decrease of administrative work by connecting systems containing metadata with one another**. With respect to the administrative systems of funding agencies, an OKB could update relevant records with scholarly communications ingested from open infrastructures (e.g., CrossRef) which aggregate metadata from the journals as sources. As a result,

researchers need not manually enter metadata into systems such as ISAAC themselves.[48] Likewise, an OKB decreases the need to enter metadata into institutional CRIS systems where this metadata is already available in open infrastructures or already entered by other researchers in their respective institutional CRIS.

2.2.2 Library & IT

For the library & IT user group we identified the main potential benefit of an OKB as the ability to share metadata between institutes. As shown in Figure 3, **an OKB proposes to establish a feedback loop between institutional CRISs and the OKB** where metadata entered or updated by one institute is available to all. Furthermore, metadata is ingested from open infrastructures such as ROR (Research Organization Registry) [49] or ISNI (International Standard Name Identifier)[50] which provides readily available metadata on organisations that a research institute may collaborate with that can be ingested in CRISs rather than manually entered.

2.2.3 National science policy

For national science policy we identified three main potential benefits of an OKB. First, that the metadata in systems such as ISAAC is improved in quality and coverage. As described above with respect to researchers, an OKB supports this by connecting metadata from journals and from institutional CRISs with ISAAC.

Second, an OKB may render visible differences in interpretations of what counts as a publication and aid harmonising the record of scholarly communications. Since an OKB proposes a feedback loop between institutional CRISs and the OKB as discussed with respect to library & IT, **an OKB renders visible how different institutes maintain the content in their respective CRISs**.

Finally, by aggregating the metadata on scholarly communications in an open infrastructure, **an OKB facilitates the development of additional research intelligence services which may report more reliable, transparent and relevant (next generation) metrics**. Note here too that an OKB facilitates a feedback loop between the data layer and the research intelligence services. If a service enriches the metadata with additional information, for example with SDG classifications, this can be fed back to an OKB (and subsequently to institutional CRISs).

2.2.4 Institutional policy

For institutional policy we identified two main benefits of an OKB. First, that **the quality and coverage of metadata in institutional CRISs is improved by establishing a feedback loop between institutional CRISs and the OKB**. As a result, SEP and KUOZ reports will be improved and potentially require less administrative work.

Second, by **enabling additional research intelligence services on top of metadata from multiple institutional CRISs**, an OKB enables better positioning of research institutes in the landscape of Dutch research institutes. Furthermore, institutional policy

---

[48] This is not to say that systems such as ISAAC no longer need a form for input. It is imaginable that use cases remain where researchers want to manually enter metadata, for example in the case of papers that will be published soon.

[49] [ror.org]

[50] [isni.org]

makers may develop additional (next generation) metrics that facilitate the analysis of institutional strategies.

2.2.5 Metadata preservation

Finally, by improving the data quality and coverage in institutional CRISs, the **metadata quality for subsequent infrastructures in the metadata chain improves**. An OKB might additionally feed directly into NARCIS, leading to a smoother harvesting of metadata for NARCIS. It is furthermore imaginable that **in the future an OKB could supersede NARCIS**[51] as an infrastructure with sustainable funding and governance.

[51] In their recent strategic programme, DANS has not included NARCIS. In an interview about the programme the director of DANS Henk Wals states that NARCIS is expected to be made redundant by new initiatives. [dans.knaw.nl]

3 Possible characteristics of an OKB

In this chapter we explore possible characteristics of an open knowledge base to identify more clearly what an OKB may look like. In section 3.1 we explore dimensions underlying an OKB (governance, critical mass, technical architecture, data scope, data quality, international connection, service development and commercial engagement, finances and funding). In section 3.2 we identify three possible models that emerge from these dimensions (API-standards, Warehouse, Research Environment). Finally, in section 0 we explore four scenarios for developing the Warehouse model (current situation, Elsevier solution, bespoke solution, parallel pilots).

3.1 Dimensions underlying an OKB

3.1.1 Governance

Main considerations for governance

- Interviewees agreed governance should be completely public.
- Networked governance with a central team in collaboration with a network of stakeholders seems desirable.
- It is essential that the responsible team has a clear mandate to make strategic decisions.
- A responsible team (e.g., working group or taskforce or otherwise) within existing organisational structures (SURF/VSNU) appears the most feasible form of governance.
- 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, GDPR, 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 and how agreement should be reached on strategic decisions. Furthermore, an important question is how an infrastructure can be achieved that improves upon limitations of the current situation, with more or less the same stakeholders. In our analysis of governance models, we distinguish several distinct questions: 1) public-private collaboration, 2) mandate formation, 3) organisation.

Public-private collaboration

The first question with respect to governance is: to what extent is the inclusion of **public or private parties** desirable? The current situation (where universities individually license commercial software) 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.[52] 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

[52] [elsevier.com]

or that the software is licensed to be open source. Interviewees emphasised that this final model would make an Open Knowledge Base open infrastructure 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 Dutch public sector, this was argued to represent a stronger case since it starts from the international connection (see §3.1.6).

Mandate formation

The second question with respect to governance is: how is a mandate formed to make strategic decisions and request or demand commitment from stakeholders? Central to this 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.[53] 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. 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. 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 metadata that is provided by research institutes and have no mandate to provide feedback on the scope or quality of that metadata. As a result, several interviewees noted that the metadata in these systems lacks in quality and therefore has limited practical value. An opportunity

[53] This risk finds precedence in the failure in 2010 of a Dutch consortium led by SURF to agree on a uniform CRIS system under the title NL-RIS, after which Dutch institutes each individually had to license CRIS systems; most ended up choosing Pure.

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).

A **middle ground is to organise networked governance** consisting of a central team with experts from research institutes and representatives at participating institutes. Since metadata is largely sourced from institutional CRIS systems, participating institutes should allocate personnel as points of contact for the central OKB operatives.

Legal entity organisation

The third question regarding governance is: how should governance be organised? The central team should receive an official mandate to make central decisions and should receive allocated time (up to full-time) to work on an OKB.[54] Depending on the model chosen for an OKB (see section 3.2 below), configurations are possible where a central team maintains a system, receiving input from working groups consisting of experts employed at research institutes. To formalise their mandate, the centralised team should be embedded in a recognisable legal entity. From the interviews and Dialogic's experience with similar projects, we see three possible models of legal entities.

1. A (non-profit) company or foundation with research institutes as clients.
2. A cooperative with research institutes as members.
3. A team embedded within existing structures.

When establishing OKB governance as a **company or foundation** this offers the advantage of being able to make strategic decisions relatively easily. Research institutes do not have a direct say in daily operations of the company but take a client relation where they can join or leave depending on whether the company's strategy suits their needs. As such, this presents a fully top-down governance model. In contrast, when establishing OKB governance as a **cooperative**, research institutes are owners of the legal entity and have a direct say in strategic decisions. Examples of such cooperatives in the Netherlands are Academic Transfer[55] and SURF[56]. This presents a bottom-up governance model, although the director or board of the cooperative may be given a certain level of autonomy.

However, these two models present two complexities that negatively affect feasibility. First, the added **complexity of establishing a new legal entity**, rather than the adoption of existing organisational structures. For example, how to sustainably finance the new legal entity is a complex matter. Especially in the case of a cooperative, a question is what advantage this poses when such a cooperative already exists with SURF. A counterpoint to this argument may be that the OKB governance proposal aims to achieve different results in a landscape without changing the actors in that landscape. Such legal entities furthermore require a higher level of overhead in the form of a CEO, fully employed staff and support staff. Second, in these two models the services offered by the legal entity should be licensed through competitions due to the **tendering obligation**, a legal requirement for public institutes such as universities[57]. This means that after the establishment of the legal entity, it is still possible that the services are not licensed after losing the tender competition.

---

[54] An example of such a networked yet centralised 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]

[55] About AcademicTransfer [academictransfer.com]

[56] The SURF cooperative [surf.nl]

[57] Aanbestedingsplicht [pianoo.nl]

The third option of a **responsible team** embedded within existing structures therefore appears to be the most feasible. During development of an OKB this responsible team may organise itself as a project team, working group or taskforce, followed by an operational team once an OKB becomes operational. Examples of such teams are the Dutch taskforce on Responsible Management and Research Information and Data of the VSNU, or the SURF programme *Acceleration plan educational innovation with ICT*[58]. Such teams place fewer demands on establishment, finances and overhead compared to legal entities. When this team is internal to the SURF, the **tendering obligation** is furthermore shifted from the individual research institutes to the cooperative.[59] This means that when an OKB is developed in-house under SURF, there is no tendering obligation.[60] Furthermore, if OKB development is outsourced, this tendering can be conducted singularly by SURF, rather than by each participating institute individually.

Based on interviews and comparisons with other systems, we estimate that the responsible 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, 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 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)

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 clear the benefits for early adopters.

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

[58] Over Versnellingsplan [versnellingsplan.nl]

[59] Algemene lidmaatschapvoorwaarden cooperatie SURF U.A., article 4 [surf.nl]

[60] Commentaar op Aanbestedingswet 2012 art. 2.24a (Aanbestedingsrecht) (2016) [sdu.nl]

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.

Furthermore, on a technological level, critical mass is readily available since Metis and Pure offer CERIF API-endpoints, and Pure can already ingest metadata from other systems as necessary for a metadata feedback loop.

3.1.3 Technical architecture

Main considerations for technical architecture

- Interviewees noted the opportunities for a centralised rather than federated architecture.
- Federated architecture is the most feasible in the short-term.
- Federated architecture lacks technical scalability and may lead to performance issues.
- A centralised architecture may be recommendable to establish a metadata feedback loop.

The central question for the technical architecture is whether to follow a **federated or centralised approach**. In a federated approach, an OKB is a connected infrastructure of 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 will 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 subsequently decide on the amount (and types) of data disclosed, providing further enrichments or disclosing a minimal set of data. A federated model thus has the advantage of **organisational 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. Finally, a federated architecture is readily within reach since Metis and Pure offer CERIF API-endpoints, making this approach feasible in the short-term.

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)[61] then implementation and sustainability is hampered by the lack of expertise with LOD approaches at the institutional level.

[61] [lod-cloud.net]

Alternatively, in a centralised architecture all data is collected in a single, centralised, system. This makes analysis easily scalable, since all data is accessible from a single point, in contrast with a federated architecture. Another advantage of a centralised system that was mentioned in several interviews is that data quality can be more easily harmonised. 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 outside 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).[62] As such, **to enable a feedback loop of metadata between institutional CRISs and the OKB (**as represented in Figure 3 in section Positioning of an OKB in the landscape), **it may be recommendable to choose a centralised architecture**.[63]

A possible disadvantage of centralisation is that the development of a centralised architecture is less flexible, since it is more difficult to add different data fields later on. As such this increases the risk of path-dependency, where design choices early in the process determine possible research questions in the future.[64] Another 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 a centralised 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 important, but rather the quality of the data (content) and the user-friendliness of the services that run on top of the infrastructure.

3.1.4 Data scope

Main considerations for data scope

- Most interviewees saw metadata as most feasible, rather than only identifiers or full texts of publications.
    - Abstracts would be desirable to be included as part of metadata, but depends on copyright.
    - Research datasets are out of scope.
- Interviewees noted the opportunity for an OKB to emphasise non-traditional research output in contrast with existing bibliometric systems.
- Metadata on funding grants and projects (linked to researchers and publications) would be highly valuable but may not be readily available from existing systems.
- Fine-grained personal data facilitates use cases but adds complexities related to privacy and GDPR.

The data scope is closely aligned to which use cases are deemed desirable. As such, the chosen data scope should facilitate the use cases which an OKB should minimally facilitate. We distinguish three separate considerations with respect to scope, namely scope of 1) scholarly communications, 2) (meta)data, and 3) metadata on researchers.

---

[62] [ma-graph.org]

[63] The alternative is that all participating institutional CRISs compare and share metadata with one another, which may prove to be more complex following Metcalfe's law.

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

The discussion of data scope has two questions that need to be considered. First, what metadata is desired or required to support use cases? Second, what metadata can actually be collected from metadata providers such as institutional CRIS systems?

Scope of scholarly communications

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**. In this sense, the primary data scope is comparable to the current scope of NARCIS.

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.[65] 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.

A consideration with regard to scope of scholarly communications is with respect to **non-traditional research output** such as datasets, software, scholarly communications aimed at the public, or educational material such as Open Educational Resources (OER). Such data should be included to align with next generation metrics, including those related to 'Erkennen & Waarderen'[66]. Several interviewees noted such output can often already be registered in institutional CRIS systems but that these are not aggregated on the national level due to lack of consensus and fear of lacking data quality (notably, that output overviews become flooded with blog posts rather than publications).

This extension of the traditional data scope of publications to include both funding grants as well as non-traditional research output may however demand investments from institutes and funding agencies to expose this metadata. Several interviewees noted that NWO, the main funder of science in the Netherlands, currently does not provide an API. Investments in an OKB (see also §3.1.8 below) may therefore demand investments in other locations of the infrastructure as well.

Scope of (meta)data

Interviewees disagreed about how much data should be collected about publications, non-traditional research output and grants. The options that were considered are to collect 1) **only identifiers**, 2) **metadata**, or 3) **full texts of publications**.

In the minimal model, an OKB contains only identifiers and relations between those identifiers. In this case, an OKB 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.

---

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

[66] VSNU et al. (2019). Ruimte voor ieders talent.

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.[67]

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 §**Error! Reference source not found.** above). However, abstracts are not clearly part of open metadata. The Initiative for Open Abstracts (I4OA) estimates that 6.6% of all works with a Crossref DOI and 8.3% of journal articles disclosed their abstracts via Crossref.[68] One interviewee noted that under the Taverne Amendment, abstracts can be made available via institutional repositories.[69] According to I4OA, abstracts can be retrieved for academic use of abstracts, but due to copyright may not always be republished.[70] An OKB could then at least provide identifiers as part of metadata to retrieve abstracts from institutional repositories, without republishing abstracts.

Finally, in the most complete model, an OKB contains the identifiers and metadata as well as the full texts of publications, non-traditional research outputs 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, large publishers have a (big) lead since they already have large numbers of full texts at their disposal.

It should furthermore be noted that we while we consider metadata on research datasets to be within scope of an OKB, the **research datasets themselves are explicitly out of scope** of an OKB. Reasons are that, in contrast with metadata on scholarly communications, research datasets may be very large (up to several terabytes), highly dynamic (up to multiple

---

[67] 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.

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

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

[70] "To clarify the situation from a legal standpoint, copyright directives permit free academic use of abstracts, for example for text mining and fact extraction purposes, without the need to obtain separate permission from the publisher. However, since the abstract is a creative work protected by copyright, third parties are not permitted to republish the text of an abstract unless the license under which it is originally published permits such republication." Initiative for Open Abstracts (2020). [i4oa.org], consulted 19 January 2021.

updates per day, e.g., satellite photos) and very diverse (ranging from satellite data to videos to scans of ancient manuscripts). For such datasets there is moreover no (immediate) benefit of harmonizing from multiple repositories, in contrast with an integrated overview of scholarly communications metadata. Such use cases for research data are instead better served by infrastructures such as DANS EASY, EOSC[71] and other open access repositories.

Scope of metadata on researchers

Finally, a third consideration is the extent to which an OKB should contain data on scholars themselves. Besides input (funding grants) and output (publications and other works) an interesting addition to an OKB could be the **projects** and project members who work on the funded research.[72] While some institutional CRIS systems contain metadata on projects, several interviewees noted that for other institutes project metadata is embedded in HR systems, which may make it more difficult to ingest this metadata in an OKB.

Furthermore, 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).[73] 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. Considering such complexities, fine-grained HR data is likely undesirable at least in the short-term. Instead, an OKB could include HR data later on, or could provide an API to connect open metadata with closed HR metadata from institutional systems.

3.1.5 Data quality

Main considerations for data quality

- Data quality deemed essential.
- Data quality and commitment are part of a reinforcing cycle.
- Improving data quality can occur at institutional and/or central level.
- Clear inclusion of provenance of data is necessary to sustain data quality.
- Data quality is a necessary outcome rather than a prerequisite of a functioning OKB.

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. Interviewees noted three consequences of low-quality data endanger the utility of the data. First, it prevents the development of compelling services. For example, an open access monitor that can only reliably report the data from two institutes but not others does not provide a compelling business case. Second, it prevents the development of trust in a data set. This means that even if a party were to develop a service (e.g., an open access monitor), lack of trust in the underlying data means there is a lack of trust in the output of the service. Finally, lack of

[71] [eosc-portal.eu]

[72] For an example of a NARCIS experiment in which this was undertaken see the project page *Language in Interaction* [narcis.nl]

[73] 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]

adoption of the data leads to a lack of commitment to the infrastructure, leading to a vicious cycle of low-quality data input and output.

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. However, several interviewees noted that since this data is collected from the institutional CRIS systems, the **low-quality of the data is a problem of commitment from institutes** rather than technical problem. It is for this reason that an OKB demands a strong governance with a mandate to request data improvement from participating institutes.

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. An OKB could then **establish a continuously improving feedback loop** where metadata is improved at the central level, fed back into CRIS systems, resulting in CRIS systems providing better metadata to an OKB. Finally, 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

- Interviewees agreed a national OKB is feasible and useful.
- It is necessary for an OKB to be positioned in and interoperable with the international landscape of infrastructures on scholarly communications.
- 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)[74], 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 nowadays use ORCID identifiers.[75] By following international standards, a Dutch OKB could in the future be connected to international infrastructures, rather than risk isolation. Several interviewees noted that an OKB should therefore adopt the **Common European Research Information Format (CERIF)**[76] for exchange of metadata between services (both as input and as output). One interviewee noted that this metadata format is readily supported by CRIS-systems such as Pure and Metis for output and by other scholarly infrastructures such as OpenAIRE for input. In short, while the initiative for an OKB can be sustained from a national context, it is **absolutely necessary for an OKB to be interoperable in the international landscape** of infrastructures on scholarly communications, particularly OpenAIRE.

A **limitation of starting at a national scale will 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 assessment, 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.
- It is 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 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 or research institutes to develop tools and services, leading to the risk that ultimately the data sits unused (see before, §3.1.3). The use cases described

[74] [wiki.surfnet.nl]

[75] NARCIS supports most internationally standardised identifiers and maps DAI identifiers with ORCID identifiers.

[76] EuroCRIS (n.d.) Main features of CERIF. [http://eurocris.org/]

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 private 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** for two reasons. First, to prevent that the chosen tools and services become entangled with the data infrastructure, reducing the flexibility of the data infrastructure and preventing unforeseen use cases, tools and services in the future. Second, 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 private 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'[77] in the public sector).

Considering this disagreement on the inclusion of services, it is advisable to launch a number of **minimal services** for use cases that demonstrate the utility of the data and provide a compelling example for commitment. The decision whether to include more sophisticated services should be postponed until the data infrastructure is in place.

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[78] (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[79] (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. A downside to this approach is that it may lead to "attribution stacking"[80] where subsequent services and services building on top of services require increasingly restrictive licenses.

3.1.8 Finances and funding

Main considerations for finances and funding

- The costs of an OKB strongly depend on the model chosen.

---

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

[78] CC0 "No rights reserved" [creativecommons.org]

[79] About The Licenses [creativecommons.org]

[80] Mozilla Science (n.d.) License stacking. [mozillascience.github.io]

- 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 with respect to finances and funding are: what would an OKB would cost? And who should cover these costs? This dimension largely depends on decisions made with respect to the dimensions discussed in the previous paragraphs. 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.

Finally, an OKB may necessitate investments at other places in the national scholarly metadata infrastructure. For example, the inclusion of funding grants or projects may demand investments from respectively funding agencies and research institutes in local systems to disclose this relevant metadata, as noted above in §3.1.4).

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 OC&W (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. These **models should not be seen as alternatives, but rather as sequential models with increasing extensions of scope** in technological complexity and functionality. Below we introduce the three models. See Table 1 for an overview of the three models and how they are positioned on each dimension.

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	Exists with Metis and Pure CERIF API-endpoints	At least five participating institutes + I	At least five participating institutes + I
Technical architecture	Federated	Centralised	Centralised
Data scope	Identifiers, metadata, possibly full texts	Identifiers, metadata	Identifiers, metadata
Data quality	Quality assured by institutes	Quality assured and possibly enriched by central entity and feedback loop	Quality assured and possibly enriched by central entity and feedback loop
International connection	International standards	International standards, national warehouse	International standards, national environment
Service development	None (possibly minimal demonstrators)	None (pure data layer) (possibly minimal demonstrators)	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) Local institutional annual costs: <€0.5 million (personnel)	Start-up costs: €2-3 million (software development and hardware acquisition) Annual costs: €2 million (mainly personnel) Local institutional annual costs: <€0.5 million (personnel)
Advantage	Can be achieved relatively quickly and cheaply, critical mass exists in CERIF API-endpoints	Data quality can be harmonised and enriched.	Data is immediately usable demonstrating utility of the data
Risks	Ends up not truly "open" with limitations in APIs	More difficult to request additional data, thereby increased path-dependency	Services need strong usability focus and sufficient flexibility for multiple user groups

The first model is what we call the **Open Knowledge API-standards**. In this model, an OKB is merely a set of standards and guidelines 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). Furthermore, the API-standards model is readily within reach since Metis and Pure both already offer CERIF API-endpoints. 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, and that it is insufficient to establish a feedback loop.

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*, 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) *centralised*, which provides the advantage to harmonise data quality and to centralise 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 separate 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. A centralised OKB may introduce a narrower data scope than local institutional CRIS systems, while a federated OKB could offset such scoping to the local API points. For example, one interviewee mentioned that while some universities include master theses in their CRIS systems, others do not. On the other hand, a centralised Warehouse offers the advantage of integration 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 in which the data within the scope of an OKB can be consulted, analysed and possibly visualised. This model extends the previous models in the addition of services that allow user interaction with the data. Governance is *public*, insofar as the research environment 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) *centralised*, which enhances performance of the research environment. An advantage of this model is that it provides reference services that demonstrate the utility of the data stored in an OKB and provides references for the development of alternative metrics. For example, the research environment 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 research environment could eventually replace existing commercial services for a number of use cases and thereby lead to cost savings.

Here too, the Open Knowledge Research Environment model should be considered an **extension of the Warehouse model** rather than a replacement; it essentially proposes the Warehouse model but with services on top.

3.2.1 Support for OKB models

Interviewees noted that **the API-standards model is insufficient to sustain an OKB**. 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. *This model, however, provides the Warehouse model with data, and should consequently be pursued anyhow*.

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 and institutional commitment**. Without tools and services that demonstrate the utility of the data, there is a risk the data sits unused by end users. However, the **Warehouse model is sufficiently beneficial for establishing a feedback loop** to improve metadata quality and coverage. Yet several interviewees noted that **the Research Environment model is desirable to sustain an OKB**. 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. This final aspect **may lead to an additional feedback loop between the Warehouse and services** to expand or further improve metadata quality and coverage.

In conclusion, the Warehouse model is largely preferred by interviewees and facilitates the feedback loop to improve metadata quality and coverage.

3.3 Scenarios for implementing an Open Knowledge Warehouse

The dimensions described above suggest that a bespoke Open Knowledge Warehouse, designed and developed specifically for the Dutch public sphere, may be preferential. However, to make an informed decision about the Warehouse model in this section we compare four scenarios: 0) maintain the current situation, 1) adopt the Elsevier solution, 2) create a bespoke solution, 3) run parallel pilots.

3.3.1 The 0-scenario: maintain the current situation

As described above in §2.1.5, NARCIS harvests institutional CRISs in a centralised system and deduplicates objects. To a large extent, NARCIS thereby already provides an Open Knowledge Warehouse that is publicly governed (based at DANS) and contains open metadata that can be reused without restrictions (with the exception of organisations and persons due to GDPR concerns)[81].

[81] Terms of Use [narcis.nl]

However, as mentioned above, A limitation of the current situation that was mentioned in interviews is that NARCIS can only collect the metadata that is provided by research institutes and has no mandate or method to provide feedback on the scope or quality of that metadata. As a result, several interviewees noted that the metadata in these systems lacks in quality and therefore has limited practical value. **Establishing a feedback loop therefore does not appear to be feasible without significant adjustments** on both levels of technology and governance.

Furthermore, interviewees noted there is a lack of commitment to NARCIS from research institutes. Instead, research institutes may even prefer to use commercial citation databases to gain overviews of scholarship. The 0-scenario thereby **does not address the core concerns** of academic independence or data coverage and quality. Furthermore, in their recent strategic programme, DANS has not included NARCIS. In an interview about the programme the director of DANS Henk Wals[82] states that NARCIS is expected to be made redundant by new initiatives.

3.3.2 The 1-scenario: license a commercial product

The first alternative to the current situation is to license a commercial product. In this scenario a market study and/or tender process is conducted to identify which commercial products satisfies the requirements of an OKB best.

As an example of this scenario, as part of the VSNU agreement with Elsevier[83] a proposal is available to implement the Pure Community Module[84]. This module enables the integration and deduplication of metadata from institutional CRISs (both Pure as well as others). Although a feedback loop is currently not available, this is on the roadmap.

Based on the guidelines as introduced in section 1.1, licensing a commercial product may satisfy conditions for an OKB related to the metadata remaining open, unrestricted access, provenance and open standards (in the case of the Pure Communit Module; CERIF, OpenAIRE and Dublin Core standards). As such, the 1-scenario addresses the core concern of data coverage and quality on the technological level. This scenario thereby satisfies the OKB as a technological proposal, but likely not as a governance proposal. While public OKB governance may place requests for adjustments to metadata scope (i.e., non-traditional scholarly output) or to the development roadmap, eventually the decision and responsibility for the commercial product lies with private enterprise. For many commercial product, the software is not open source and not all algorithms in the content pipeline are visible and transparent (but may be indicated through provenance).

In short, **the 1-scenario offers a feasible, usable and affordable solution that can be implemented relatively quickly**. However, this offers no guarantees beyond the duration of agreements or beyond the functionality already offered. As such, this scenario **does not adequately address concerns of academic independence** insofar as this concerns independent strategic decision making and governance on functionality and scope.

---

[82] Focus op FAIR (2020). DANS. [dans.knaw.nl]

[83] Open science platform products and services agreement, p. 103 [vsnu.nl]

[84] Pure Community Module [elsevier.com]

3.3.3 The 2-scenario: develop a bespoke solution

The second alternative concerns the development of a bespoke solution. This would consist of building the technology to harvest, store and enrich metadata[85], as well as the technology for establishing the feedback loop. Since it is developed based on the requirements and wishes of the OKB governance, this solution could in principle address all concerns on a technological level. This need not entail that all components of an OKB need to be built from the ground up, insofar as open source components may be reused from existing infrastructures such as NARCIS, OpenAIRE or the Curtin Open Knowledge Initiative. The collection of code and software can be open source and algorithms can be made transparent at least on the level of code and documentation.

Furthermore, since OKB governance is fully in control of metadata scope and development roadmap, this solution addresses academic independence insofar as the public sphere owns and controls the storage and distribution of metadata. Note that this does not necessarily prohibits private enterprises from eventually developing the OKB, which could be organised through a tender process. A bespoke solution demands strong governance however to make clear strategic decisions and protect the scope of the OKB.

In short, compared to the 1-scenario, a bespoke solution may (on the short-term) be **less usable and more costly**. It does however **fully address the core concerns** of academic independence and data quality and coverage.

3.3.4 The 3-scenario: parallel maintaining, licensing and developing

The above three scenarios each present advantages and disadvantages with respect to resources, costs and addressing the core concerns of academic independence and data quality and coverage. The 3-scenario could be to run parallel pilots that adopt the advantages of each scenario and critically assess the limitations offered by the disadvantages. This includes licensing a commercial product to assess to what extent this indeed results in improved data quality and coverage, as well as to assess to what extent requests for metadata scope (i.e., non-traditional scholarly output) and development roadmap are not satisfied. In parallel, a baseline OKB could be developed to assess the technological complexity of bespoke development. Insofar as the data in the commercial product is openly available, the bespoke development can reuse this data in parallel. Furthermore, during this process the current situation could be maintained to fully assess to what extent the licensed or developed systems improve upon the current situation. After sufficient time to experience advantages and disadvantages of each scenario, an **evaluation could be conducted whether to continue and possibly expand the 1-scenario or whether this fails to meet demands, necessitating the 2-scenario**. Moreover, if through the pilots it is discovered that the feedback loop does not lead to improvements in data quality and coverage, **it could be decided to maintain the 0-scenario**.

[85] This may reuse technology available within NARCIS.

4 Roadmap for OKB development

In this chapter we sketch a roadmap for developing an OKB. In section 4.1 we briefly discuss to what extent development can be conducted in parallel or serially. In section 4.2 we present an overview of the different phases and discuss each phase with respect to tasks and deliverables.

4.1 Parallel or serial development

The three models presented in the previous chapter do not represent mutually exclusive options but are instead sequential models where a decision should be reached whether to pursue more advanced models. The *Research Environment* model depends on the *Warehouse* model, which in turn depends on the *API-standards* model. A decision can be made whether to stop at a more basic model or pursue more advanced models. Deciding this early on allows development to be done in parallel, speeding up the process. However, several interviewees noted that **to secure feasibility of an OKB**, it is advisable to **take one (small) step at a time to prevent design and development from stalling due to discussions about future aspects**. For example, the development of standards model should not await the building of consensus about the inclusion of services and whether or not to pursue the Research Environment model. It is, however, important to keep sight of such future considerations with respect to building the appropriate governance and technical support.

We distinguish between several phases, starting from the current phase. Each phase should end with a decision on the governance and finance model for the next phase.

4.2 Overview of phases

In the following sections we describe the different phases as represented schematically in Figure 4. Phases are numbered chronologically, where phases with the same number can be conducted in parallel. Per phase we identify the main deliverables. Phase 2B 'Design Warehouse' (elaborated in §4.5.1) is the phase in which the scenario's identified in section 0 are decided upon.

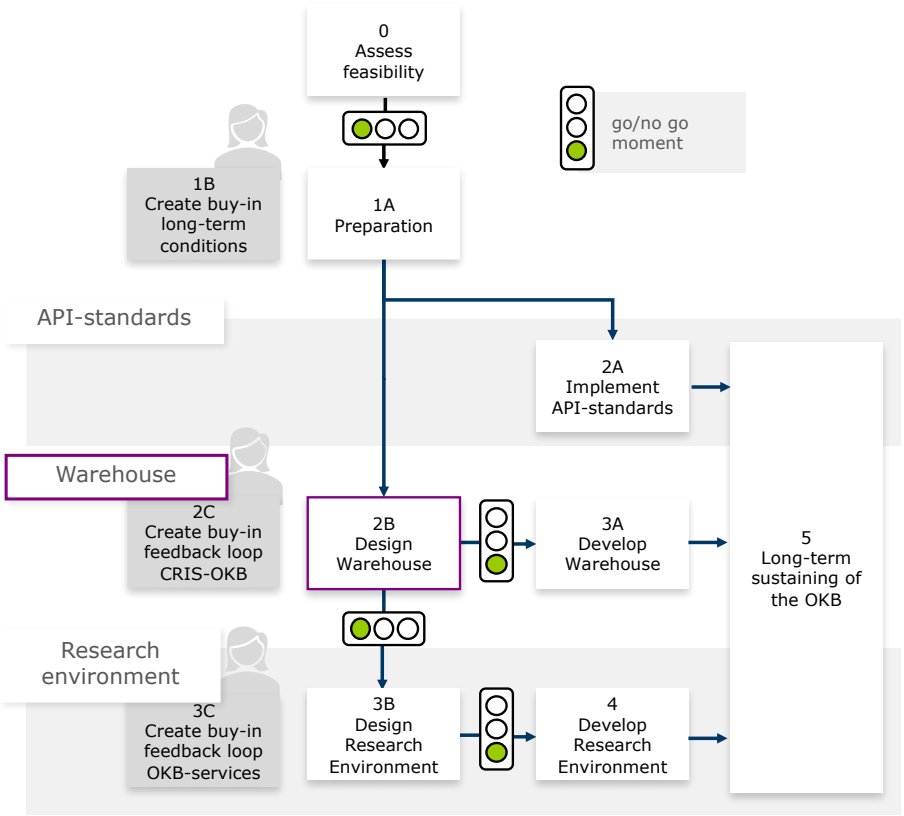


Figure 4. Overview of roadmap. Blue boxes represent phases of design and development. Green boxes represent necessary buy-in from stakeholders to continue. Red arrows represent go/no go decision points. Phase 2B 'Design Warehouse' is highlighted in purple to indicate the importance of this phase where a decision is made between the scenarios discussed in section §0.

4.3 Phases related to preparation

4.3.1 Phase 0: Assess feasibility (current phase)

Main deliverables of the current phase

1. Feasibility study
2. Decision on desirability
3. Governance and finance model next phase

The feasibility phase is the current phase. The current report can be considered as the first deliverable, assessing the **feasibility** of an OKB.

The second deliverable should be the **go/no go** on pursuing an OKB. If it is decided that an OKB should not be pursued the roadmap effectively ends with this phase.

If it is decided that an OKB is indeed desired, the third deliverable should be a formalised model for governance and financing of the next phase, notably the responsible team that will oversee the OKB.

4.3.2 Phase 1A: Preparation

Main deliverables

1. Establishment of responsible team
2. Identification of conditions for long-term sustainability

3. Creation of buy-in for long-term conditions
4. Governance and finance model next phase

If an OKB is indeed pursued, the preparation phase serves to build the business case for an OKB and to create buy-in for long-term conditions of an OKB.

The first deliverable should be the **establishment of the responsible team** that will oversee the OKB, as described in the previous phase.

The second deliverable should be the **identification of conditions for long-term sustainability** of an OKB. In this deliverable, the long-term prospects of the OKB, explored with the current feasibility study (chapter 2), should be formalised. Furthermore, this deliverable should identify the necessary expertise for developing and sustaining an OKB and explore to what extent this expertise is readily available within public institutes or demands outsourcing to private enterprises. Building on the Guiding Principles on Management of Research Information and Data[86], this deliverable should then suggest to what extent an OKB should be developed and owned publicly or what aspects may be outsourced. We suggest that in the case of outsourcing, this deliverable explores to what extent expertise can be developed within the public sphere so as not to become dependent on private enterprises for sustainability. That is, the conditions for long-term sustainability should give insights into exit strategies when contracts with private enterprises (e.g., the current VSNU deal with Elsevier) end. This deliverable is preparatory to the long-term phase (§4.7.1) with which the roadmap ends and provides the basis for further discussions on conditions. It should therefore be seen as a 'living document' that is continually updated.

Finally, the third deliverable is to model governance and financing of the next phases 2A and 2B. These two parallel phases may largely consist of the same governance and financing but are described separately as they might demand different expertise and pace of work.

4.3.3 Phase 1B: Create buy-in for long-term conditions

Main deliverables
1. Critical mass of participating research institutes

Parallel to phase 1A in which the conditions for long-term sustainability of an OKB are identified, phase 1B encompasses **the creation of buy-in for said conditions for long-term sustainability**.

This encompasses that a **critical mass of participating research institutes** (see also §3.1.2) is consulted and commitment is gained for long-term sustainability of an OKB.

4.4 Phases related to the API-standards model

4.4.1 Phase 2A: Implement API-standards

Main deliverables of the establishment phase
1. Inventory of national infrastructure and API endpoints
2. Harmonised documentation of API endpoints
3. Governance and finance model(s) next phase

[86] VSNU (2020). Guiding Principles on Management of Research Information and Data.

The API-standards OKB is developed in the API-standards phase. It is in this phase that the OKB becomes part of the national infrastructure. The phase ends with the API-standards model in place.

The first deliverable is to create an **inventory of the (inter)national infrastructure and API endpoints**. This includes to what extent institutional CRIS systems have appropriate metadata output for an OKB.[87] Furthermore, an inventory of open infrastructures that contain relevant metadata for enrichment of an OKB (e.g., CrossRef, Orcid, ROR, etc.) should be undertaken. Finally, this includes the extent to which the desired data scope demands additional investments at other places in the national infrastructure, notably the interoperability with ISAAC at NWO and the ability (on a technological level) of institutional CRISs to ingest data from an OKB for local data enrichment.

The second deliverable builds upon this inventory. This harmonised **documentation of API endpoints** provides a central viewpoint on where metadata on scholarly communications can be found and accessed. The difference with the earlier deliverable is that while the earlier report is a description of what API endpoints are available, this deliverable should go a step further and provide developers with the information to build services that make use of these API-endpoints.

The contents of this deliverable are furthermore **dependent on the outcomes of phases 2B and 2C** which run parallel to this phase. That is, if in phase 2B it is decided not to pursue a centralised Warehouse model, and in phase 2C buy-in is created for a feedback loop between institutional CRISs and the OKB, then this deliverable should describe how the API-standards model will sustain this feedback loop.

Finally, with the API-standards model thus implemented and documented, the phase ends with a model for governance and financing of the long-term phase.

4.5 Phases related to the Warehouse model

4.5.1 Phase 2B: Design Warehouse

Main deliverables

1. Requirements for Warehouse
2. Market exploration
3. Decision for in-house or outsourced development
4. Decision on Warehouse model desirability and development
5. Governance and finance model next phase

In this phase it is explored how to make metadata on scholarly communications available through a centralised data warehouse. This phase works toward the decision whether or not to pursue the Warehouse model. This phase can therefore run parallel to phase 2A.

In the first deliverable, the **requirements for the Warehouse** are explored and agreed upon and formalised. This deliverable builds upon the long-term conditions identified in phase 1A. These requirements consider both what is desired of the Warehouse model, but also the requirements in expertise and resources to sustain the Warehouse model. Furthermore, this deliverable should describe how the Warehouse may ingest metadata enrichments and improvements from tools and services.

---

[87] Several interviewees noted that both Pure and Metis already offer CERIF API-endpoints.

Second, a **market exploration** should be conducted to identify what products and infrastructures exist that may inspire or contribute to an OKB. During this process, both private enterprises as open infrastructures should be explored to consider how they might facilitate (components of) an OKB. We advise that this exploration is conducted in parallel to the setting of requirements in the first deliverable so that requirements are informed by experiences of existing systems. However, for clarity and separation of concerns we consider these two separate deliverables.

Based on the first and second deliverables, the third deliverable should be to a **decision whether to develop the Warehouse in-house** (i.e., by a team of developers employed at a public institute part of the OKB governance) **or by outsourcing** to a private enterprise. First, this considers the **decision between the scenarios** described in section 0; licensing a commercial product or developing a bespoke solution. Second, in case of a bespoke solution, it considers whether to develop in-house or outsource development through a tender. This deliverable should specify which route is taken and what the financial consequences are. Furthermore, in case it is decided to develop in-house it should be specified where this team of developers could be based. Furthermore, if it is decided to outsource development due to lacking expertise or resources in the public sphere, the long-term phase model (§4.7.1) should be updated to include a roadmap for gaining the required expertise and resources so that the sustainability of the OKB does not become dependent on private enterprises.

The fourth deliverable should be the **go/no go** on pursuing the Warehouse model. If it is decided that an OKB should not be pursued the roadmap with respect to the Warehouse effectively ends with this phase.

If it is decided to pursue the Warehouse model, the final deliverable is the governance and finance model for the next phase of Warehouse development. This model should take into account the decision on whether to outsource or develop the Warehouse in-house.

4.5.2 Phase 2C: Create buy-in for feedback loop CRIS-OKB

Main deliverable
1. Creation of buy-in for CRIS-OKB feedback loop

As argued in chapter 2, one of the main benefits of an OKB is to establish a feedback loop in order to improve data quality and coverage. Parallel to phase 2A and 2B, phase 2C encompasses the **creation of buy-in for a feedback loop between institutional CRISs and the OKB**. This encompasses that a **critical mass of participating research institutes** (see also §3.1.2) is consulted and commitment is gained for implementing, using and sustaining this feedback loop. What this feedback loop will entail on a technological level is dependent on the decision whether to pursue the centralised Warehouse model (phase 2B) and the implementation of the API-standards model (phase 2A).

4.5.3 Phase 3A: Develop Warehouse

Main deliverables
1. Hiring of personnel or private party for development
2. Warehouse
3. Governance and finance model(s) next phase

In the development phase the Open Knowledge Warehouse is developed according to the requirements and means decided upon in the previous phase.

The first deliverable is **process**-oriented and consists of putting the first deliverable into action. Namely, in this step personnel are hired (for in-house development) and embedded in an organizational structure or the tender is published to hire a private party for development. The tender should be based on the requirements analysis of phase 2B.

The second deliverable then is the **Warehouse**. This may consist of multiple sub-deliverables and components.

This phase ends when the Warehouse is implemented. Development then moves from start-up to operations, maintenance and continued development. The final deliverable should be the governance and finance model for sustaining the Warehouse and running operations. The governance model for the long-term phase (§4.7.1) should therefore be updated to include the Warehouse.

4.6 Phases related to the Research Environment model

4.6.1 Phase 3B: Design Research Environment

Main deliverables

1. Requirements for Research Environment
2. Decision for in-house or outsourced development
3. Decision on Research Environment model desirability and development
4. Governance and finance model next phase

In this phase it is explored how to make metadata on scholarly communications available for analysis through a Research Environment. This phase works toward the decision whether or not to pursue the Research Environment model. This phase can therefore run parallel to phase 3A.

In the first deliverable, the **requirements for the Research Environment** are explored and agreed upon and formalised. This deliverable builds upon the long-term conditions identified in phase 1A. These requirements consider both what is desired of the Research Environment model, but also the requirements in expertise and resources to sustain the Research Environment model. Furthermore, since this model includes tools and services, this deliverable should describe **use cases** that should be facilitated.

Based on the first deliverable, the second deliverable should be to a **decision whether to develop the Research Environment in-house** (i.e., by a team of developers employed at a public institute part of the OKB governance) **or by outsourcing** to a private enterprise. This deliverable should specify which route is taken and what the financial consequences are. Furthermore, in case it is decided to develop in-house it should be specified where this team of developers could be based. Note that this decision need not be the same as the decision for the development of the Warehouse model. By separating the data layer and the services layer, it is possible (and perhaps recommendable) that development is separated so to prevent vertical integration and vendor lock-in. Furthermore, if it is decided to outsource development due to lacking expertise or resources in the public sphere, the long-term phase model (§4.7.1) should be updated to include a roadmap for gaining the required expertise and resources so that the sustainability of the OKB does not become dependent on private enterprises.

The third deliverable should be the **go/no go** on whether or not to pursue the Research Environment model. If it is decided that an OKB should not be pursued the roadmap with respect to the Research Environment effectively ends with this phase.

If it is decided to pursue the Research Environment model, the final deliverable is the governance and finance model for the next phase of Research Environment development. This model should take into account the decision on whether to outsource or develop the Research Environment in-house.

4.6.2 Phase 3C: Create buy-in for feedback loop OKB-services

Main deliverable
1. Creation of buy-in for CRIS-OKB feedback loop

As argued in chapter 2, one of the main benefits of an OKB is to establish a feedback loop in order to improve data quality and coverage. The services and tools that process the metadata in the OKB may enrich or improve this metadata, which can be fed back into the OKB. Parallel to phase 3A and 3B, phase 3C encompasses the **creation of buy-in for a feedback loop between institutional CRISs and the OKB**. This encompasses that a **critical mass of participating research institutes** (see also §3.1.2) is consulted and commitment is gained for implementing, using and sustaining this feedback loop. What this feedback loop will entail on a technological level is dependent on the design of the Warehouse model (phase 2B) and the decision whether to pursue the Research Environment model (phase 3B).

4.6.3 Phase 4: Develop Research Environment

Main deliverables
1. Hiring of personnel or private party for development
2. Research environment
3. Governance and finance model next phase

In the development phase the Open Knowledge Research Environment is developed according to the requirements and means decided upon in the previous phase.

The first deliverable is **process**-oriented and consists of putting the first deliverable into action. Namely, in this step personnel are hired (for in-house development) and embedded in an organizational structure or the tender is published to hire a private party for development. The tender should be based on the requirements analysis of phase 3B.

The second deliverable then is the **Research Environment**. This may consist of multiple sub-deliverables and components.

This phase ends when the Research Environment is implemented to support the identified use cases. Development then moves from start-up to operations, maintenance and continued development. The final deliverable should be the governance and finance model for sustaining the Research Environment and running operations. The governance model for the long-term phase (§4.7.1) should therefore be updated to include the Research Environment.

4.7 Phase related to long-term sustainability

4.7.1 Phase 5: Long-term sustaining of the OKB

Main tasks
1. Establishment long-term governance and finances
2. Sustaining of open knowledge base
3. Sustaining buy-in
4. Roadmap acquiring and sustaining expertise and resources

The long-term phase is where short-term start-up moves into long-term maintenance. This phase therefore includes no deliverables which would indicate an end of the phase, instead we suggest several tasks which should continuously be maintained in this phase.

First, long-term governance and finances should be established. The responsible team established in phase 1A thereby moves toward a **sustainable organisation**. Rather than a project team overseeing initial development, the team becomes an operational team managing, maintaining and continuing development of the OKB.

Second, long-term **sustaining of the open knowledge base**, depending on the deliverables and decisions in phases 2A, 2B, 3A and 3B.

Third, long-term **sustaining of buy-in** as created in phases 1B, 2C and 3C. This may include creating buy-in beyond the critical mass. Furthermore, as noted in section 1.2, the feedback loop may demand renewed discussions about how to define what counts as scholarly output. This includes the positioning of the OKB, where the OKB organically grows to become part of the national landscape.

Finally, in case it is decided in phases 2B and 3A that expertise or resources are currently lacking for in-house development, a **roadmap** should be designed and followed to build said **expertise and resources**. This is necessary to prevent long-term vendor lock-in, where the sustainability of the open knowledge base becomes dependent on the expertise and resources of the private party to which development was outsourced. For example, it is possible that Elsevier will develop the Open Knowledge Warehouse as part of the current VSNU agreement with Elsevier (see also §3.3.2). This roadmap should ensure that upon renewal of the agreement the Dutch research institutes have acquired a position of **potential competition** so that contract renewal is not from a position of dependence.[88] As such, it is possible that no go and/or outsourcing decisions made in phases 2B and 3B are retaken in the future.

[88] As also suggested by the Chief Innovation Officer of SURF. SURF loopt niet alleen aan de hand van grote tech-spelers (2021). *ScienceGuide* [sciencegeuide.nl]

5 Conclusions and recommendations

In this chapter we present out conclusions and recommendations. In section 5.1 we list the main conclusions from the feasibility study. In 5.2 we list seven recommendations to initiate and sustain an OKB.

5.1 Conclusions

5.1.1 Positioning and purpose of an Open Knowledge Base

1. The OKB proposal should be considered as both a **technological proposal of a data layer and metadata feedback loop** as well as a **governance proposal of establishing buy-in** to sustain a metadata lifecycle
2. The core values of an OKB can be summarised as two concerns. First, to **protect academic independence** by opening up the metadata and metrics underlying assessments of scholarship. Second, to **improve and enhance the quality and coverage of metadata** available in the Dutch landscape of infrastructures on scholarly communications.
3. An OKB may lead to **several improvements for user groups** with respect to the current landscape of infrastructures on scholarly communications:
    a. For **researchers** an OKB can **decrease the amount of administrative work** by connecting systems that possess metadata with systems that require metadata.
    b. For **library & IT** an OKB can decrease the amount of work by **sharing metadata between institutional CRISs**.
    c. For **national science policy** an OKB can 1) lead to **better quality and coverage metadata** in administrative systems, 2) create a **better and richer overview of scholarship** produced by grants and 3) **enable the creation of next generation metrics** that better align with national priorities and strategies.
    d. For **institutional policy** an OKB can 1) **improve the quality and coverage of metadata** in institutional CRISs necessary for **SEP and KUOZ reports** and 2) create a better and richer overview of national scholarship in which institutes are positioned.
    e. An additional benefit may be that an OKB could improve the quality and coverage of metadata in NARCIS or in the future (gracefully) supersede NARCIS as an infrastructure with sustainable funding and governance.

5.1.2 Dimensions underlying an OKB

4. Governance designed as **networked governance** with a central working group in collaboration with a network of stakeholders is preferred.
    a. A **central working group** (or taskforce or otherwise) within existing organisational structures (SURF/VSNU) appears the most feasible form of a responsible team.
    b. It is essential that the central working group has a **clear mandate to make strategic decisions**.

    c. Governance requires approximately five to fifteen people (depending on the expertise and skills needed), which need **dedicated time and resources** for their tasks.
5. An OKB **need not start with all institutes** but should identify a **critical mass** to establish a minimum viable product.
6. Stakeholders are **divided about the choice between federated and centralised architecture**.
    a. A federated architecture is the most feasible in the short-term.
    b. A federated architecture may lack technical scalability and may lead to performance issues.
    c. A centralised architecture may moreover be preferred to establish a metadata feedback loop.
7. The **data scope should include metadata of traditional objects** such as publications, grants, authors, institutes, funding agencies, **as well as metadata of non-traditional research output** such as datasets, software, scholarly communications aimed at the public and open educational resources.
    a. The data scope should be chosen to provide a compelling **alternative overview to existing (commercial) citation databases**.
    b. The data should include the metadata on all objects as well as identifiers.
    c. Further investigation is necessary to explore to what extent abstracts or full texts could be incorporated.
    d. To improve the links between authors, grants and scholarly output, it may be **recommendable to include projects** as objects. These are however currently not usually available in public data.
    e. More fine-grained HR data to provide overviews of Dutch academia should be explored in the future but is not feasible in the short-term.
8. **Improving data quality and coverage should be a main concern** of an OKB.
    a. Data can be improved both at the institutional CRIS level as well as at the central level through **integration with open infrastructures** (CrossRef, Orcid, ROR, etc.)
9. An OKB is most feasible when **starting at the national level**.
    a. A national OKB should however **follow international data standards** (e.g., CERIF, but also international identifiers).
    b. A national OKB may however lack certain use cases such as international benchmarks.
10. Development may be done by private enterprises
    a. It is essential to **separate development of the data layer and the tools & services layer.**
    b. Interviewees **disagreed whether tools & services should be part of an OKB**.
11. The costs of an OKB strongly depend on the chosen model
    a. A **federated architecture is relatively easily achievable**, since most API-endpoints already exist.
    b. 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.
    c. **Governance should prevent early adopters from bearing disproportionate weight** in funding during start-up.
12. Based on the above dimensions we identify three possible and feasible models. These models are not alternatives but as sequential models with increasing extensions of scope
    a. The **API-standards model** consists of a set of standards and guidelines of metadata that each institute or organisation should provide through an openly

available API. This model is readily within reach, since Metis and Pure (used by most research institutes) offer API-endpoints using the CERIF metadata standard. 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, and that it is insufficient to establish a feedback loop.

b. The **Warehouse model** consists of a centralised data warehouse where metadata is collected from the API-endpoints, deduplicated and harmonised. Metadata can furthermore be enriched and expanded from other open infrastructures (e.g., CrossRef, Orcid, ROR).
c. The **Research Environment** expands the warehouse with the addition of research intelligence services and tools that **demonstrate the utility** of the data stored in the OKB and provides **references** for the development of alternative metrics. Some interviewees argued that such services are necessary to attract user engagement and institutional commitment. Furthermore, such services and tools may establish an **additional feedback loop** between the Warehouse and the services to expand or further improve metadata quality and coverage.
d. We conclude that the **Warehouse model is most desirable and feasible**.

13. With respect to implementation of the Warehouse model, we identify four possible scenarios.
    a. 0-scenario: **maintain the current situation**. This scenario offers **open metadata in an open infrastructure** (i.e., NARCIS) but lacks a feedback loop or governance model to gain commitment to improve data quality and coverage. Research institutes subsequently remain dependent on commercial citation databases to gain overviews of scholarship. The 0-scenario thereby **does not sufficiently address the core concerns** of academic independence or data coverage and quality.
    b. 1-scenario: **license a commercial product**. This scenario offers **open metadata in a closed infrastructure** that is feasible, usable and affordable and can be implemented relatively quickly. It is possible that metadata quality and coverage is improved through a feedback loop. The infrastructure is however dependent on a commercial offering. As such, the 1-scenario **does not adequately address academic independence**.
    c. 2-scenario: **develop a bespoke solution**. This scenario builds an OKB by developing or reusing (parts of) open source software to offer **open metadata in an open infrastructure** that **establishes a feedback loop and governance to gain commitment**. Compared to the 1-scenario, a bespoke solution may (on the short-term) be **less usable and more costly**. It does however **fully address the core concerns** of academic independence and data quality and coverage.
    d. 3-scenario: **parallel maintaining, licensing and developing**. In this scenario **all three scenarios are pursued** for an agreed time period. In the final year an evaluation is conducted to **assess the impact of the scenarios** on the improvement of metadata quality and coverage. Furthermore, scenarios 1 and 2 can be compared to assess to what extent requests for metadata scope and development roadmap are satisfied or not and to assess technological complexity of bespoke development.

5.1.3 Roadmap

14. For the OKB as technological proposal the roadmap identifies phases to design and develop all three models, with separate go/no go points.

15. For the OKB as governance proposal the roadmap identifies the need to create buy-in at three points
    a. Long-term conditions
    b. Feedback loop CRIS-OKB
    c. Feedback loop OKB-services
16. Long-term sustainability of an OKB requires four tasks
    a. A **sustainable organisation** with dedicated time and resources.
    b. **Sustaining of the technology** underlying the OKB.
    c. **Sustaining of buy-in** and creating buy-in beyond the initial critical mass, **organically positioning the OKB** in the national landscape, and **renewed discussions about definitions what counts** as scholarly output.
    d. Roadmap to develop necessary expertise and resources to sustain **potential competition** with private enterprises so that contract renewals are not from a position of dependence.

5.2 Recommendations

1. **Pursue (at minimum) the Warehouse model** to connect existing infrastructures in the Netherlands by collecting, storing, enriching and distributing metadata. As such an OKB addresses the concerns of academic independence and metadata quality and coverage.
2. **Collect in the Warehouse metadata on traditional objects** such as publications, grants, authors, institutes, funding agencies, **as well as non-traditional research output** such as datasets, software, scholarly communications aimed at the public and open educational resources. Include moreover **projects** to connect these objects in time.
3. **Establish a responsible team with a clear and strong mandate and dedicated time and resources** to make and pursue strategic decisions. This is possible by forming a working group within SURF, which subsequently provides legal basis for in-house development or tenders.
4. **Attract strong leadership** to lead the responsible team. To create buy-in, governance requires leadership that can **negotiate with the top level of research institutes** (rectorate and/or institutional policy). Prevent discussions and negotiations about an OKB to be limited to the library & IT level.
5. **Establish buy-in for metadata feedback loops** between institutional CRISs and the OKB (continuous metadata enrichment and enhancement) as well as between the OKB and research intelligence services (algorithmic enrichment).
6. **Aim at improving rather than replacing currently available systems** in the landscape of infrastructures on scholarly communications. This positioning should grow organically over time. With respect to NARCIS **it should be explored to what extent an OKB may (gracefully) supersede NARCIS as an infrastructure with sustainable funding and governance**.
7. **Initiate an OKB from the national level** but **position it in and ensure interoperability with the international landscape** of infrastructures on scholarly communications **by following international standards and data models**.
8. **Identify the necessary expertise and resources** to sustain an OKB and **create a roadmap to develop these conditions in the public sphere**. Even in case an OKB is developed by private enterprises or based on off-the-shelf products, it should remain possible to offer a **potentially competitive scenario** so that contract agreements and renewals are not from a position of dependence.

Appendix 1. Interviewees

#	User group	Name	Affliation
1.	Library & IT	Alastair Dunning	TU Delft
2.	Library & IT	Arjan Schalken	UKBsis
3.	Library & IT	Armand Guicherit	TU Delft
4.	Library & IT	Corno Vromans	Tilburg University
5.	Library & IT	Ed Simons	Radboud University Nijmegen, euroCRIS
6.	Library & IT	Enno Meijers	KB National Library of the Netherlands
7.	Library & IT	Erik Flikkenschild	Leiden University Medical Center
8.	Library & IT	Erna Sattler	Leiden University
9.	Library & IT	Hans Schoonbrood	Radboud University Nijmegen
10.	Library & IT	Hanna-Mari Puuska	CSC
11.	Library & IT	Henk Wals	DANS
12.	Library & IT	Herbert van de Sompel	DANS
13.	Library & IT	Hylke Annema	Tilburg University
14.	Library & IT	Lambert Heller	Technische Informationsbibliothek Hannover
15.	Library & IT	Maarten Steenhuis	KB National Library of the Netherlands
16.	Library & IT	Magchiel Bijsterbosch	SURF
17.	Library & IT	Marc van den Berg	KB National Library of the Netherlands
18.	Library & IT	Martijn Kleppe	KB National Library of the Netherlands
19.	Library & IT	Nick Veenstra	TU Eindhoven
20.	Library & IT	Sören Auer	Technische Informationsbibliothek Hannover
21.	Institutional policy	Bianca Kramer	Utrecht University
22.	Institutional policy	Hans Ouwersloot	Maastricht University, DAIR
23.	Institutional policy	Jeroen Bosman	Utrecht University
24.	Institutional policy	Karin Maex	University of Amsterdam
25.	Institutional policy	Maurice Vanderfeesten	VU University Amsterdam
26.	National science policy	Darco Jansen	VSNU
27.	National science policy	Hans de Jonge	NWO
28.	National science policy	John Doove	SURF

#	User group	Name	Affliation
29.	Researchers	Cameron Neylon	Curtin University, COKI
30.	Researchers	Egon Willighagen	Maastricht University
31.	Researchers	Jason Maassen	eScience Center
32.	Researchers	Kaitlin Thaney	Invest in Open Infrastructure
33.	Researchers	Ludo Waltman	Leiden University, CWTS
34.	Researchers	Natalia Manola	University of Athens, OpenAIRE
35.	Researchers	Paul Wouters	Leiden University
36.	Researchers	Rob van Nieuwpoort	eScience Center
37.	Researchers	Rudi Bekkers	TU Eindhoven
38.	Researchers	Sarah de Rijcke	Leiden University, CWTS
39.	Researchers	Wilco Hazeleger	Utrecht University
40.	Private enterprise	Bo Alroe	Dimensions
41.	Private enterprise	Joep Verheggen	Elsevier
42.	Private enterprise	Max Dumoulin	Elsevier
43.	Private enterprise	Theo Pillay	Elsevier
44.	Private enterprise	Tijmen Altena	IDfuse

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