Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

Roland Bertelmann[1 I], Julia Boltze[2 II], Klaus Ceynowa[3 III IV], Jürgen Christof[4 V], Katja Faensen[5 I], Matthias Gross[IV], Cornelia Hoffmann[VI], Thorsten Koch[6 II], Monika Kuberek[7 V], Stefan Lohrum[II], Heinz Pampel[8 I], Markus Putnings[9 VI], Regina Retter[III], Beate Rusch[10 II], Hildegard Schäffler[III], Konstanze Söllner[11 VI], Ronald Steffen[12 V], Eike Wannick[13 VII]

DeepGreen: Open-Access-Transformation in der Informationsinfrastruktur – Anforderungen und Empfehlungen Version 1.0

[1] 0000-0002-5588-0290 [2] 0000-0002-0819-4271 [3] 0000-0002-8257-8070 [4] 0000-0001-8642-0425
[5] 0000-0002-0091-9637 [6] 0000-0002-1967-0077 [7] 0000-0002-1672-5271 [8] 0000-0003-3334-2771
[9] 0000-0002-6014-9048 [10] 0000-0001-7664-4097 [11] 0000-0002-6263-7846 [12] 0000-0002-1001-4188
[13] 0000-0002-4723-4555

[I] Helmholtz-Gemeinschaft, Helmholtz Open Science Office
[II] Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV)
[III] Bayerische Staatsbibliothek (BSB)
[IV] Bibliotheksverbund Bayern (BVB)
[V] Technische Universität Berlin, Universitätsbibliothek (TU Berlin)
[VI] Friedrich-Alexander-Universität Erlangen-Nürnberg, Universitätsbibliothek (FAU)
[VII] Deutsches Rundfunkarchiv Potsdam, vormals: Helmholtz-Gemeinschaft, Helmholtz Open Science Office

ZIB Report 21-03 (Februar 2021)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30 84185-0
Telefax: +49 30 84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Inhaltsverzeichnis

Abstract

DeepGreen ist ein Service, der es teilnehmenden institutionellen Open-Access-Repositorien, Open-Access-Fachrepositorien und Forschungsinformationssystemen erleichtert, für sie relevante Verlagspublikationen in zyklischer Abfolge mithilfe von Schnittstellen Open Access zur Verfügung zu stellen. Die entsprechende Bandbreite an Relationen zwischen den Akteuren, diverse lizenzrechtliche Rahmenbedingungen sowie technische Anforderungen gestalten das Thema komplex. Ziel dieser Handreichung ist es, neben all diesen Themen, die begleitend beleuchtet werden, im Besonderen Empfehlungen für die reibungslose Nutzung der Datenübertragung zu liefern. Außerdem werden mithilfe einer vorangestellten Workflow-Evaluierung Unterschiede und Besonderheiten in den Arbeitsschritten bei institutionellen Open-Access-Repositorien und Open-Access-Fachrepositorien aufgezeigt und ebenfalls mit Empfehlungen angereichert.

1 Der Dienst DeepGreen

Als Projekt im Rahmen der Ausschreibung „Open-Access-Transformation“ (DFG, 2014) der Deutschen Forschungsgemeinschaft (DFG) wurde DeepGreen in zwei Projektphasen von 2016 bis 2021 zu einer zentralen Informationsinfrastruktur zur Förderung der Open-Access-Transformation entwickelt. Als Datendrehscheibe stellt der Dienst DeepGreen einen rechtssicheren Workflow zur Verarbeitung von automatisierten Datenlieferungen von wissenschaftlichen Verlagen an institutionelle Open-Access-Repositorien, an Open-Access-Fachrepositorien und Forschungsinformationssysteme sicher.

Die Ziele und erreichten Meilensteine sind in zahlreichen Publikationen des DeepGreen-Projektteams nachvollziehbar (DeepGreen, 2020a) und wurden in zwei Projektphasen von den Projektpartnern Bayerische Staatsbibliothek (BSB), Bibliotheksverbund Bayern (BVB), Universitätsbibliothek der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Helmholtz-Gemeinschaft – Helmholtz Open Science Office, Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV) und Universitätsbibliothek der Technischen Universität Berlin (TU Berlin) realisiert.

Auf Software-Grundlage des „Jisc Publications Router“ wurde die DeepGreen-Datendrehscheibe weiterentwickelt, um Verlagsdatenlieferungen in Abhängigkeit von rechtlichen Rahmenbedingungen automatisiert an die berechtigten Repositorien mit Metadaten und Volltext verteilen zu können.

Der Dienst kann aktuell folgende Lizenz-, Vertrags- und Publikationsvarianten unterstützen:

- Allianz- und Nationallizenzen
- FID-Lizenzen
- Gold Open Access
- Transformationsverträge

Die Funktionsweise der Datendrehscheibe konnte im Verlauf der erweiterten Testphase seit Herbst 2019 mit 60 teilnehmenden institutionellen Open-Access-Repositorien und sechs angeschlossenen Forschungsinformationssystemen sowie durch drei Fachrepositorien mit zahlreichen Nutzungsfällen durchgespielt werden.

Zunächst ist für die Nutzung der DeepGreen-Datendrehscheibe die Erstellung eines Kontos für die jeweilige Institution bzw. das jeweilige Fachrepositorium nötig. Mit den vom DeepGreen-Team mitgeteilten Zugangsdaten können sich die Repositoriumsbeauftragten auf der Weboberfläche einloggen und weitere Einstellungen zur Einrichtung des Kontos vornehmen. Für einen erfolgreichen Betrieb ist dabei vor allem die Erstellung einer Affiliationsdatei bzw. ISSN-Journalliste notwendig, durch die erst die Zuordnung der Artikel erfolgen kann. Das Konto wird mit den Angaben zur Repositorien-Software ergänzt.[1] Wird die SWORD-Schnittstelle für eine automatisierte Übernahme der Artikel ins Repositorium

[1] Siehe dazu auch die FAQ-Liste unter https://deepgreen.kobv.de/de/erweiterte-testphase-des-deepgreen-projektes/faq-haeufig-gestellte-fragen-zum-erweiterten-test/

verwendet (vgl. *3.2.3 SWORD*), müssen auch hier die entsprechenden Einträge vorgenommen werden.

Es werden von DeepGreen selbst keine Artikel gespeichert und veröffentlicht, die Artikeldaten werden lediglich an die entsprechenden Institutionen weitergeleitet. Die bei den einzelnen Artikeln hinterlegten Metadaten werden ausgeliefert, wie sie vom Verlag bereitgestellt werden.[2] DeepGreen verändert die Verlagsdaten nicht. Somit hat der Dienst keinen Einfluss auf die Qualität der Verlagsmetadaten. Die rechtliche Verantwortung rund um die Veröffentlichung der Publikationen ist in schriftlichen Vereinbarungen geregelt (vgl. *2.2.1 DeepGreen – Institution*). So ist beispielsweise im Hinblick auf die Beachtung einer möglichen Embargofrist die jeweilige veröffentlichende Institution selbst verantwortlich. Eine Dublettenerkennung ist aktuell nicht möglich. Die Integration einer Filterfunktion ist derzeit in Planung.

Die folgenden Erläuterungen beleuchten unterschiedliche Aspekte der Arbeit mit DeepGreen, angefangen von den rechtlichen Rahmenbedingungen über technische Spezifikationen zu einzelnen Repositorientypen hin zu Workflowempfehlungen. Außerdem werden Wege zu zukünftigen Weiterentwicklungen thematisiert.

[2] DeepGreen wandelt die von den Verlagen gelieferten Metadatenformate (meistens JATS/XML) in für Repositorien handhabbare Formate um und stellt diese über verschiedene Schnittstellen bereit.
An der SWORD-Schnittstelle werden, je nach Repositorien-Software, verschiedene Ausgabeformate zur Verfügung gestellt: https://oa-deepgreen.github.io/user_docs/repository.html, s. Abschnitt 4.1

2 Rechtliche Rahmenbedingungen

Anliegen von DeepGreen ist es, die rechtliche Prüfung von Zweitveröffentlichungen zu erleichtern. Entscheidend für die rechtliche Prüfung sind einerseits die Konditionen, die unterschiedliche Lizenztypen für die Zweitveröffentlichung bieten und andererseits die vertraglichen Beziehungen, die DeepGreen für seinen Service eingeht.

2.1 Lizenztypen

DeepGreen verteilt Publikationen, die im Rahmen unterschiedlicher Lizenztypen erstveröffentlicht wurden. Damit gehen Konsequenzen für die Möglichkeiten zur Zugänglichmachung in einem Open-Access-Repositorium einher. Im Folgenden werden die wichtigsten Lizenztypen für DeepGreen vorgestellt.

2.1.1 Allianz- und Nationallizenzen

Ausgangspunkt für die Entwicklung des Dienstes DeepGreen sind Allianz- und Nationallizenzen. Die seit 2011 mit Teilförderung der DFG angebotenen Allianz-Lizenzen, die Bibliotheken per Opt-In die Möglichkeit bieten, einen lesenden Zugriff auf laufende Jahrgänge aktueller Zeitschriften zu erhalten, beinhalten eine spezielle Open-Access-Klausel, die weitreichende Zweitveröffentlichungsrechte vorsieht. Die Grundform dieser Klausel lautet:

> *„Autoren aus autorisierten Einrichtungen sind ohne Mehrkosten berechtigt, ihre in den lizenzierten Zeitschriften erschienenen Artikel in der Regel in der durch den Verlag publizierten Form (z.B. PDF) zeitnah in institutionelle oder disziplinspezifische Repositorien ihrer Wahl einzupflegen und im Open Access zugänglich zu machen. Das gleiche Recht besitzen die autorisierten Einrichtungen, denen die jeweiligen Autoren angehören.“* (DFG, 2015, S. 9; Verbundzentrale des GBV, 2019)

Daraus ergeben sich vier Charakteristika:

- Die Version, die Open Access zugänglich gemacht werden kann, ist in der Regel das Verlags-PDF, also die sogenannte Version of Record.
- Berechtigt sind die Autor*innen und/oder die Institutionen, denen sie angehören. Es gibt keine Einschränkung auf Erstautor*innen oder Corresponding Authors, alle Autor*innen sind gleichermaßen berechtigt.
- Die Zweitveröffentlichung kann entweder in einem institutionellen oder disziplinspezifischen Open-Access-Repositorium nach Wahl erfolgen.
- Es gilt eine Embargo-Frist.

Im Einzelfall können die Regelungen in den Lizenzverträgen von dieser Standard-Formulierung abweichen.

Die Geltung dieser Open-Access-Klausel erstreckt sich teilweise auch auf Nationallizenzen, die ältere Zeitschriften-Jahrgänge mit voller DFG-Förderung für alle deutschen

Forschungseinrichtungen ohne zusätzliche Kosten zugänglich machen. Die Publikationen der Allianz-Lizenzen gehen nach Ablauf einer festgelegten Zeitspanne, der Moving Wall, in Nationallizenzen über. Darüber hinaus bestehen Nationallizenzen, die nicht an Allianz-Lizenzen gekoppelt sind.

Ob Open-Access-Rechte aus Allianz-Lizenzen über die Moving Wall automatisch auf den weiteren Teilnehmerkreis der Nationallizenz übergehen, ist nicht pauschal für alle Lizenzen zu beantworten. Alle an DeepGreen beteiligten Verlage, die sowohl Allianz- als auch Nationallizenz-Teilnehmer beliefern, stimmen der Interpretation zu, dass mit dem Übergang von Inhalten aus der Allianz- in die Nationallizenz alle Nationallizenz-Teilnehmer die Open-Access-Klauseln aus den Allianz-Lizenzen wahrnehmen dürfen.[3]

Welche konkreten Bedingungen für welche Lizenzverträge gelten, wird derzeit von den Verhandlungsführer*innen an zwei Stellen zentral nachgewiesen. Einerseits gibt es die Open-Access-Angaben in der Elektronischen Zeitschriftenbibliothek (nachfolgend EZB), die über die Open-Access-EZB-Schnittstelle ausgegeben werden können (EZB, 2016). Andererseits wird auf der Webseite der Nationallizenzen eine tabellarische Übersicht der Open-Access-Rechte von Allianz- und Nationallizenzen angeboten (Verbundzentrale des GBV, 2019). Mit Hilfe dieser beiden Werkzeuge kann die einzelne Institution die Korrektheit der Lieferungen von DeepGreen überprüfen.[4]

Manche Allianz-Lizenzen gehen nach Ende der DFG-Förderung in Nationalkonsortien über, für die weiterhin die ursprünglichen Open-Access-Rechte gelten. In diesen Fällen pflegen die Verhandlungsführer*innen die Informationen in der EZB und der Tabelle für die Open-Access-Rechte der Nationallizenzen weiter.

Inhalte der Allianz- und Nationallizenzen stehen üblicherweise nicht unter einer Creative-Commons-Lizenz. Für den korrekten Nachweis der Lizenz in den Repositorien empfiehlt die Handreichung „Open-Access-Rechte in Allianz- und Nationallizenzen“ daher die folgende Formulierung:

> *„Dieser Beitrag ist mit Zustimmung des Rechteinhabers aufgrund einer (DFG-geförderten) Allianz- bzw. Nationallizenz frei zugänglich.“*
>
> *„This publication is with permission of the rights owner freely accessible due to an Alliance licence and a national licence (funded by the DFG, German Research Foundation) respectively.”* (Stöber, 2012, S. 11)

Institutionelle Open-Access-Repositorien dürfen Inhalte der Allianz- und Nationallizenzen dank der Standard-Klausel ohne die explizite Erlaubnis der Autor*innen veröffentlichen. Ungeachtet dessen empfiehlt es sich als Serviceleistung der Bibliotheken, die betroffenen Autor*innen über die Zweitveröffentlichung zu informieren. Für disziplinspezifische Open-Access-Repositorien sieht die Klausel kein vergleichbares Recht vor. Zwar dürfen die Inhalte auch in diesen Repositorien veröffentlicht werden, das Recht dazu haben jedoch nur der

[3] DeGruyter ist hier ausgenommen. DeGruyter beliefert via DeepGreen derzeit nur Teilnehmer der Allianz-Lizenz, die Nationallizenz beinhaltet keine gesonderten Green-Open-Access-Rechte.
[4] Die Angaben auf diesen Seiten werden nach bestem Wissen von den Verhandlungsführer*innen gepflegt, sind jedoch nicht rechtsverbindlich. Rechtlich verbindlich ist immer der Wortlaut des Vertrags.

jeweilige Publizierende oder seine Institution. Praktisch bedeutet dies, dass disziplinspezifische Open-Access-Repositorien entweder Workflows benötigen, um die Erlaubnis der Autor*innen oder Institutionen einzuholen oder eine gesonderte Berechtigung beim Verlag einholen müssen. Ein systematischer Versuch, diese Berechtigung für Open-Access-Fachrepositorien bei den Allianz-Lizenz-Verlagen einzuholen, wurde bislang von DeepGreen noch nicht unternommen.

Nicht eindeutig geklärt ist zum Zeitpunkt der Veröffentlichung dieser Handreichung die Frage, ob ein an einer Allianz-Lizenz Teilnehmender nur an den Inhalten Zweitveröffentlichungsrechte erhält, die während der Laufzeit des eigenen Vertrags erschienen sind, oder ob der Teilnehmende auch vor seiner Beteiligung erschienene Veröffentlichungen zugänglich machen darf. Die Ausgabe der Open-Access-Rechte über die EZB legt nahe, dass ein Teilnehmender die OA-Rechte ab Beginn der Existenz der Lizenz erhält, unabhängig von seinem persönlichen Eintrittsdatum. DeepGreen spiegelt in seinen Lieferungen die Angaben in der EZB.

2.1.2 FID-Lizenzen

DFG-geförderte Lizenzen der Fachinformationsdienste für die Wissenschaft (FID) können ebenfalls eine Open-Access-Klausel beinhalten, die im Wesentlichen der Allianz-Lizenz-Klausel entspricht. Es handelt sich hierbei jedoch nicht um eine harte Anforderung der DFG, die Klausel soll nur „nach Möglichkeit“ enthalten sein (DFG, 2015, S. 8). De facto ist diese Klausel derzeit nur in den wenigsten FID-Lizenzen tatsächlich verankert.

Für FID-Lizenzen wurden drei Nutzerkreis-Modelle[5] entwickelt, die für DeepGreen unterschiedlich gut abzubilden sind. Der Community-Typ, der nur Einzelnutzenden als Angehörigen einer Fachcommunity den Zugriff auf Ressourcen erlaubt, ist für DeepGreen zum jetzigen Zeitpunkt nicht bedienbar[6]. Der Campus-Typ, bei dem sich Lizenzen auf alle Angehörige einer teilnehmenden Institution erstrecken, kann von DeepGreen wie eine Allianz-Lizenz unter Einbeziehung einer von Verhandlungsführenden gestellten Teilnehmer*innenliste verarbeitet werden. Der National-Typ, der analog zur Nationallizenz deutschlandweit allen Institutionen nach Registrierung offensteht, wird von DeepGreen wie eine klassische Nationallizenz behandelt (siehe *2.1.1 Allianz- und Nationallizenzen*).

Anders als Allianz- und Nationallizenzen werden die Open-Access-Rechte aus FID-Lizenzen nicht in zentralen Nachweissystemen wie der EZB oder der im vorherigen Unterkapitel erwähnten Tabelle auf der Nationallizenzen-Webseite vorgehalten. Die genaue Ausgestaltung der Open-Access-Rechte lässt sich nur aus den jeweiligen Verträgen ersehen. Vertragsauszüge mit Open-Access-Abschnitten sind über die jeweiligen Webseiten der FID in der Rubrik FID-Lizenzen/Produkte unter den Nutzungsbedingungen des jeweiligen Produkts einsehbar.[7]

[5] Kompetenzzentrum für Lizenzierung (2019): Lizenz- und Nutzerkreismodelle für FID-Lizenzen: Der KfL-Lizenzbaukasten, online unter https://www.fid-lizenzen.de/FIDInfo_Lizenzmodelle_20190124_v2.pdf, S. 2.
[6] Der Matchingprozess des Dienstes DeepGreen beruht auf Institutionszugehörigkeit. Ein namentlicher Abgleich von Einzelpersonen ist nicht vorgesehen.
[7] Für den FID Pharmazie beispielsweise unter https://pharmazie.fid-lizenzen.de/produkte#b_start=0

Bei Inhalten aus FID-Lizenzen ist wie bei Allianz-Lizenz-Inhalten nicht davon auszugehen, dass sie unter einer Creative-Commons-Lizenz veröffentlicht werden dürfen. Die Formulierung der Lizenzbedingungen kann in Anlehnung an die Allianz-Lizenzen erfolgen.

Bei FID-Lizenzen gilt wie bei Allianz- und Nationallizenzen: Disziplinspezifische Open-Access-Repositorien können zwar zur Zweitveröffentlichung genutzt werden, berechtigt sind hierzu jedoch nur Autor*innen und/oder deren Institutionen.

2.1.3 Gold Open Access

Die Zustellung von Publikationen, die in Open-Access-Zeitschriften erscheinen, ist für DeepGreen rechtlich einfach umsetzbar. Diese Publikationen sind bereits im Open Access erschienen und können unter Berücksichtigung der jeweiligen Lizenz – in der Regel CC BY, CC BY-NC oder CC BY-NC-ND – ohne weitere Embargos und Einschränkungen von allen Publizierenden jederzeit öffentlich zugänglich gemacht werden. So sind bei Gold-Open-Access-Publikationen disziplinspezifische und institutionelle Repositorien gleichgestellt. Die jeweilige Lizenz der Publikation ist anzugeben.

2.1.4 Transformationsverträge

Immer stärkere Bedeutung gewinnen Transformationsverträge, die es teilnehmenden Institutionen ermöglichen, neben dem lesenden Zugriff auf Subskriptionsinhalte auch Open-Access-Publikationsrechte in hybriden Zeitschriften für ihre Erstautor*innen zu sichern. Dies geht gegebenenfalls einher mit Rabattierungen auf Publikationsgebühren, die bei der Veröffentlichung in reinen Gold-Open-Access-Zeitschriften, die sich auch im Verlagsportfolio befinden, anfallen.

Publikationen aus Transformationsverträgen können demnach sowohl genuine Open-Access-Publikationen[8] sein als auch Subskriptionsinhalte, die über den grünen Weg zu bestimmten Konditionen zweitveröffentlicht werden können. Bislang bildet DeepGreen nur den genuinen Open-Access-Anteil von Transformationsverträgen ab. Hierfür werden genuine Open-Access-Publikationen von den Verlagen durch DeepGreen an die am Transformationsvertrag beteiligten Institutionen ausgeliefert. Disziplinspezifische Repositorien können, in Abhängigkeit der genauen Ausgestaltung des Verlagsvertrags, ebenfalls eine automatisierte Zustellung gemäß dem fachlich gewünschten Zuschnitt erhalten. Dieser wird durch eine ISSN-Liste der fachlich einschlägigen Zeitschriftentitel definiert.

Für die Zugänglichmachung gelten – wie auch bei Gold-Open-Access – die Bestimmungen der Lizenz, die im Regelfall die Creative Commons Lizenz „CC BY“ ist.

2.1.5 Konsortiallizenzen ohne Green-Open-Access-Komponente

Ein weiterer für DeepGreen relevanter Lizenztyp sind Publikationen aus Subskriptionszeitschriften. Hier greifen die Regelungen der jeweiligen Verlage, die in den

[8] Genuin Open Access beinhaltet Gold Open Access und Hybrid Open Access

Autor*innenverträgen niedergelegt sind und über die Webseiten der Verlage sowie – ohne Gewähr – über den Service von Sherpa Romeo[9] abgerufen werden können.

Die jeweiligen Verlagspolicies können unterschiedliche Bestimmungen an einzuhaltenden Embargos enthalten. Zum Beispiel zur nutzbaren Version – meist Pre- oder Postprint – und zum Ort der Zweitveröffentlichung: Von der persönlichen Webseite über institutionelle und fachliche Open-Access-Repositorien bis hin zu Social-Media-Netzwerken. Gemein ist den Verlagspolicies, dass allein die Autor*innen (in vielen Fällen sogar nur der Corresponding Author) zur Zweitveröffentlichung berechtigt ist. Demnach muss bei Publikationen, die gemäß der Verlagspolicy durch DeepGreen zugänglich gemacht werden, ein Vertrag mit dem jeweiligen Verlag geschlossen werden.[10]

2.1.6 Zweitveröffentlichungsrecht laut UrhG §38 (4)

Es ist möglich, dass Autor*innen aufgrund des Zweitveröffentlichungsrechts laut Urheberrechtsgesetz §38, Absatz 4 im Einzelfall weitergehende Rechte zustehen, als in Verlagspolicies oder Lizenzverträgen gewährt werden. Dieser Lizenztyp kann jedoch nicht von DeepGreen abgebildet werden[11].

2.2 Vertragliche Beziehungen

Um seinen Service zu gewährleisten, unterhält DeepGreen Verträge mit verschiedenen Partnern und greift auf bestehende vertragliche Vereinbarungen Dritter zurück. Zeichnungsberechtigt für DeepGreen ist das Zuse-Institut Berlin, an dem die Zentrale des KOBV angesiedelt ist. Nachfolgend werden die Relationen zu DeepGreen thematisiert.

2.2.1 DeepGreen – Institution

Die Beziehung zwischen DeepGreen und teilnehmenden Open-Access-Repositorien wird durch einen Haftungsausschluss definiert, der einseitig von teilnehmenden Institutionen unterzeichnet wird. Im Haftungsausschluss übernehmen die Institutionen die Verantwortung für das Zugänglichmachen der von DeepGreen gelieferten Inhalte. Dies impliziert, dass die ausgelieferten Publikationen vor Veröffentlichung auf das Vorhandensein entsprechender Rechte aus den oben genannten Lizenzen geprüft werden müssen.

2.2.2 DeepGreen – Verlag

Mit beteiligten Verlagen schließt DeepGreen Vereinbarungen. Hier werden Rechte und Pflichten der Vertragspartner vereinbart und insbesondere die Lizenzbedingungen festgelegt, die die Grundlage der automatisierten Zugänglichmachung durch DeepGreen darstellen. Im Fall von Allianz-, National- und FID-Lizenzen sowie von Transformationsverträgen wird explizit

[9] Siehe https://v2.sherpa.ac.uk/romeo/
[10] Zum Zeitpunkt der Veröffentlichung dieser Handreichung sind Auslieferungen nach Verlagspolicy noch in Vorbereitung.
[11] Hierfür fehlt schon die notwendige Datenbasis. Zudem beruht der derzeitige Verteilmechanismus auf Regeln, die für je eine gesamte Institution gelten. Abwandlungen für einzelne Forschungsprojekte oder Einzelpersonen wären mit einer massiven Komplexitätssteigerung verbunden.

auf die jeweiligen zugrundeliegenden Lizenzverträge verwiesen. Sollten Details der Open-Access-Rechte in diesen nicht abschließend geregelt sein, werden entsprechende Regelungen in die Kooperationsvereinbarung aufgenommen. Im Fall von Gold-Open-Access-Zeitschriften oder Verlagen, die auf Grundlage der eigenen Verlagspolicy kooperieren, werden die jeweiligen Lizenzbedingungen direkt in der Kooperationsvereinbarung festgelegt.

2.2.3 Institution – Verlag und Verhandlungsführer – Verlag

Bei Allianz-, National- und FID-Lizenzen sowie Transformationsverträgen sind die entscheidenden Regelungen in den Lizenzverträgen niedergelegt, die die teilnehmenden Institutionen bzw. stellvertretend die zuständigen Konsortialführer mit den Verlagen geschlossen haben. Die für DeepGreen relevanten Lizenzverträge sind ausschließlich Konsortialverträge.

DeepGreen selbst hat, im Gegensatz zu den teilnehmenden Institutionen, keinen systematischen Zugriff auf den Wortlaut der Konsortialverträge. Der Dienst DeepGreen bezieht seine Informationen aus der EZB oder aus dem direkten Austausch mit der verhandlungsführenden Institution. Rückfragen zur Auslegung einzelner Inhalte der Lizenzverträge müssen deshalb immer mit der verhandlungsführenden Institution abgeklärt werden.

3 Technische Spezifikationen

3.1 Technische Anbindung von Repositorien und Zuordnungsverfahren

Um an DeepGreen partizipieren zu können, ist es notwendig, ein Konto für die teilnehmende Einrichtung und deren Open-Access-Repositorium zu erstellen. Neben einer E-Mail-Adresse und einem Passwort wird hierfür von DeepGreen die Unterzeichnung eines Haftungsausschlusses durch die teilnehmende Einrichtung benötigt. Nach der Anmeldung ist es ratsam, dieses Konto zu kuratieren. Im Folgenden wird die technische Anbindung für institutionelle Open-Access-Repositorien dargestellt. Informationen zur Anbindung von fachlichen Open-Access-Repositorien finden sich im Unterkapitel *3.1.1 Zusatzanforderungen im Zuordnungsverfahren für Fachrepositorien*.

Der Matchingprozess der Metadaten zur korrekten Zuordnung der Publikationen für eine teilnehmende Einrichtung findet anhand der Affiliationsdatei (siehe *Tabelle 1 Beispiel einer match-config-Datei*) statt. In dieser können sämtliche zur Zuordnung relevanten Affiliationsangaben eingetragen werden. Zum Beispiel Namensvariationen der Einrichtung, E-Mail-Domains, Förderkürzel, institutionelle Identifier etc. In der Affiliationsdatei finden sich bereits von DeepGreen bereitgestellte Informationen zur Namensansetzung der teilnehmenden Institution, die auf Angaben in der Gemeinsamen Normdatei (GND) beruhen. Je ausführlicher die Affiliationsdatei von der teilnehmenden Institution kuratiert und gewartet wird, desto höher die Qualität der Datenlieferungen. Es ist daher ratsam, die Affiliationsdatei sorgfältig auszufüllen.

```
Name Variants,Domains,Grant Numbers,Dummy1,Dummy2,Keywords
Academia Fridericiana Erlangensis,,,,,
Academia Friderico Alexandrina Erlangen-Nürnberg,,,,,
Academia Friderico-Alexandrina,,,,,
Academia Regia Bavarica Friderico-Alexandrina,,,,,
Academia Regia Friderico-Alexandrina,,,,,
Bayerische Friedrich-Alexanders-Universität,,,,,
F.A.U. Erlangen-Nürnberg,,,,,
FAU Erlangen-Nürnberg,,,,,
Universitas Literarum Regia Friderico-Alexandrina,,,,,
Università di Erlangen-Nürnberg,,,,,
University of Erlangen-Nuremberg,,,,,
University of Erlangen-Nürnberg,,,,,
,fau.de,,,,
,uk-erlangen.de,,,,
,uni-erlangen.de,,,,
,,,123456-563/2,,
,,,99988/365-2,,
```

Tabelle 1: Beispiel einer match-config-Datei.

Open-Access-Repositorien, die mit den Software-Lösungen DSpace, Eprints, MyCoRe oder OPUS arbeiten, können relativ schnell und unproblematisch an die Datendrehscheibe angeschlossen werden.

Nach Anbindung des Open-Access-Repositoriums durch DeepGreen und der Einrichtung des Kontos durch die teilnehmende Einrichtung, die das Repositorium betreut, werden die Artikel

nach einem zweistufigen Matchingprozess anhand der in der Affiliationsdatei hinterlegten Informationen zugeordnet. Um festzustellen, ob die jeweilige Publikation im Rahmen eines durch DeepGreen unterstützten Lizenztypen (Allianz-, Nationallizenzen oder Nachfolgelizenzen) veröffentlicht wurde und somit von DeepGreen prozessiert werden kann, findet zunächst ein Abgleich der ISSN der Zeitschrift, in der die Publikation erschienen ist, mit einer Kollektionsliste statt, die auf den Angaben der EZB beruht. Im nächsten Schritt werden die Affiliationsangaben in den Metadaten der Publikation mit denen in der Affiliationsdatei verglichen. Bei Treffern bekommt die jeweilige Institution eine Notifikation in ihrem DeepGreen-Konto und kann den oder die Artikel automatisch mittels SWORD oder manuell weiter prozessieren. Die Prüfung der rechtlichen Grundlagen für die öffentliche Zugänglichmachung der Publikation (z.B. Einhaltung etwaiger Embargofristen oder sonstiger verlagsseitiger Einschränkungen) obliegt der Institution. DeepGreen übernimmt hier keine Haftung.

3.1.1 Zusatzanforderungen im Zuordnungsverfahren für Fachrepositorien

Da Open-Access-Fachrepositorien, im Gegensatz zu institutionellen Open-Access-Repositorien, nicht die Publikationen einer bestimmten Forschungseinrichtung sammeln, sondern den Zugang von Publikationen eines Faches im Open Access ermöglichen, kann bei der Zustellung von Publikationen von DeepGreen an Open-Access-Fachrepositorien keine Zuordnung über den Abgleich der Affiliationen erfolgen. Fachrepositorien müssen daher in der Affiliationsdatei die Spalte *„Name Variants“* auf den Wert „IGNORE-AFFILIATION“ setzen um die Zuordnung auf Basis der Affiliations-Metadaten zu deaktivieren.

Stattdessen erfolgt eine Zuordnung der von DeepGreen auszuliefernden Publikationen an ein Open-Access-Fachrepositorium über eine individuelle ISSN-Liste, welche nur diejenigen ISSN-Codes von Zeitschriften enthält, die dem Fachzuschnitt des fachlichen Open-Access-Repositoriums entsprechen. Diese Liste wird im Dialog zwischen DeepGreen und dem Fach-Repositorium erstellt. Nach individueller Absprache mit einem Verlag könnten perspektivisch auch Publikationen entsprechend weiterer Parameter von DeepGreen zugestellt werden. Hierzu könnte z.B. das Metadatenfeld *„metadata.subject“* in der Affiliationsdatei genutzt werden, damit verbunden sind jedoch Anforderungen an die Metadatenqualität der Verlage (siehe Unterkapitel *4.1.2 Besonderheiten des Workflows von Fachrepositorien*).

3.2 Schnittstellen

Da Verlage, häufig anders als Open-Access-Repositorien, mehrheitlich Metadatenformate wie CrossRef-XML, DTD NLM/NISO JATS oder ONIX (Putnings & Rusch, 2016) zur Auszeichnung von Publikationen und Metadaten im Bereich der wissenschaftlichen Zeitschriften nutzen, stellt DeepGreen eine Kompatibilität zu gängigen Repositorien-Softwareplattformen sicher. Metadaten und Volltexte der Publikationen können so automatisch in die jeweiligen Repositorien-Plattformen eingespielt werden. Um die erfolgreiche Zustellung der Publikationen an Repositorien zu unterstützen, werden folgende Schnittstellen von DeepGreen unterstützt:

- **OAI-PMH** (Open Archives Initiative Protocol for Metadata Harvesting)**:**
  Die Open-Access-Repositorien der teilnehmenden Einrichtungen können Metadaten im Dublin-Core-Metadatenformat (DC) von DeepGreen abrufen.

- **SWORD** (Simple Webservice Offering Repository Deposit)**:**
  Metadaten und Volltexte (inkl. Embargoinformation) werden von DeepGreen (per SWORD-Client) an die Open-Access-Repositorien der teilnehmenden Einrichtungen übermittelt.

- **Web-API:**
  Die Open-Access-Repositorien der teilnehmenden Einrichtungen können Metadaten und Volltexte in zwei sukzessiven Schritten via REST-API (Representational State Transfer) von DeepGreen abrufen.

3.2.1 OAI-PMH

Via Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) können die Open-Access-Repositorien der teilnehmenden Einrichtungen über HTTP-Anfragen – ohne Authentifizierung – Metadaten im Metadatenformat von DeepGreen abrufen. Die Zustellung der zu den Metadaten gehörenden Volltexte selbst kann dann nach entsprechender Authentifizierung über die Web-API oder via SWORD erfolgen.

Die OAI-PMH Schnittstelle ist über die URL

https://www.oa-deepgreen.de/oaipmh/all?verb=identify

erreichbar. Eine Liste aller von DeepGreen bereitgestellter Notifikationen, unabhängig von ihrer Zuordnung zu einer teilnehmenden Einrichtung, kann über die URL

https://www.oa-deepgreen.de/oaipmh/all?verb=ListIdentifiers&metadataPrefix=oai_dc

abgerufen werden. Die Metadaten eines einzelnen Artikels können anschließend mit dem aus der vorigen Anfrage erhaltenen *<Identifier>* als Parameter über:

```
https://www.oa-deepgreen.de/oaipmh/all
   ?verb=GetRecord&identifier=<Identifier>
   &metadataPrefix=oai_dc
```

bezogen werden. Ausführliche Informationen zur Implementierung sind in der technischen Dokumentation von DeepGreen zu finden (DeepGreen 2020b).

3.2.2 Web-API

DeepGreen stellt eine REST-API bereit, mit der Publikationen abgerufen werden können. Durch die bereitgestellten Endpunkte können die Open-Access-Repositorien der teilnehmenden Einrichtungen in einem ersten Schritt Metadaten abrufen und in einem zweiten Schritt die dazugehörigen Volltexte beziehen.

Die Basis-URL für GET- und POST-Anfragen an diese Schnittstelle lautet

https://www.oa-deepgreen.de

Eine Liste aller seit einem bestimmten Datum (hier 01.01.2020) verarbeiteten Notifikationen unabhängig ihrer Zuordnung zu spezifischen Institutionen kann z.B. über

https://www.oa-deepgreen.de/api/v1/routed?since=2020-01-01

erhalten werden. Die Metadaten einer spezifischen Notifikation können dann mittels der daraus erhaltenen *<Notification-ID>* über

https://www.oa-deepgreen.de/api/v1/notification/*<Notification-ID>*

abgerufen werden. Für den Abruf von institutionsspezifischen Notifikationslisten und Volltexten wird die Repositorien-ID benötigt bzw. der API-Key zur Authentifizierung, beides ist im DeepGreen-Account zu finden. Weitere Informationen zu verfügbaren Endpoints sind in der im Unterkapitel zuvor bereits erwähnten technischen Dokumentation (DeepGreen 2020b) recherchierbar.

3.2.3 SWORD

Die DeepGreen-Datendrehscheibe unterstützt sowohl SWORD v1 als auch SWORD v2. Zur Nutzung der SWORD-Schnittstelle durch die Open-Access-Repositorien der teilnehmenden Einrichtungen müssen folgende Punkte beachtet werden:

1. **Vorbereitung des Systems**

   a) **Aktivierung der Schnittstelle**
   Open-Access-Repositorien müssen die SWORD-Schnittstelle freischalten. Dabei sollte eine Beschreibung des Service mittels des Servicedokuments bereitgestellt werden, die auf jeden Fall eine Authentifizierung verlangt und unter einer URL abrufbar sein sollte. Des Weiteren muss ein Account für DeepGreen angelegt werden, der die Nutzung des SWORD-Services gestattet.

   b) **Collection (Sammlung)**
   Open-Access-Repositorien müssen sicherstellen, dass eine geschützte Sammlung mit einem Freischaltungsworkflow für die Aufnahme von SWORD-Importen in das System vorliegt.

   c) **Metadaten-Einspeisung**
   Je nach verwendeter Software unterstützt DeepGreen folgende Metadatenformate für den Datenabruf: DC, DTD RSC, EsciDoc, METS/MODS, NLM/NISO JATS, OPUS-XML und RIOXX. Verwendet ein Open-Access-Repositorium keines der genannten Metadatenformate ist eine Konvertierung in das systemeigene Metadatenschema nötig. Gegebenenfalls muss ein Mappingverfahren der importierten Metadaten auf interne Felder definiert werden. Hierfür empfiehlt sich die Nutzung von XSL-Transformationen.

Beispiele für Mapping-Skripte sind:

- DSpace (aus METS/MODS)[12]
- MyCoRe (aus NLM/NISOJATS)[13]

2. **Einrichtung der Zugangsdaten**
   Nach erfolgreicher Anmeldung des Open-Access-Repositoriums bei DeepGreen kann die Institution ihre SWORD-Zugangsdaten unter https://www.oa-deepgreen.de/ hinterlegen. Folgende Angaben werden benötigt:

   a) URL der geschützten Sammlung, der die Publikationen zugeordnet werden sollen.

   b) Benutzername des SWORD-Users.

   c) Password des SWORD-Users.

   d) URI des genutzten Metadataformats (Packaging preferences)

Weitere Details zu den Funktionalitäten der SWORD-Schnittstelle sind ebenfalls in der technischen Dokumentation (DeepGreen 2020b) zu finden.

[12] Siehe sword-mods-ingest.xsl: https://github.com/tuub/DSpace/blob/depositonce-6.3x/dspace/config/crosswalks/sword-mods-ingest.xsl
[13] Siehe jats2mods.xsl: https://github.com/MyCoRe-Org/mir/blob/e9336d17796df62fa6d53ba7a0001f9ac9d4a2f3/mir-module/src/main/resources/xsl/sword/jats2mods.xsl

4 Die Verarbeitung von Notifikationen auf Repositorienseite

Je nach Art des Informationssystems und den spezifischen Anforderungen der teilnehmenden Institutionen kann sich der Arbeitsablauf bei der Verarbeitung von Notifikationen unterscheiden. Spezifische Anwendungsfälle ergeben sich insbesondere durch lizenzrechtliche Anforderungen, aber auch durch die vorhandene IT-Infrastruktur.

In den folgenden Abschnitten werden verschiedene Anwendungsfälle beschrieben und Empfehlungen für einen allgemeinen Workflow für Open-Access-Repositorien zur Verarbeitung von DeepGreen-Notifikationen gegeben.

4.1 Anwendungsfälle

Basierend auf dem Feedback der teilnehmenden Open-Access-Repositorien während der Testphase konnten allgemeine Arbeitsschritte bei der Verarbeitung von Notifikationen identifiziert werden, die von verschiedenen Repositorien in vergleichbarer Weise eingesetzt werden (siehe *Abbildung 1*).

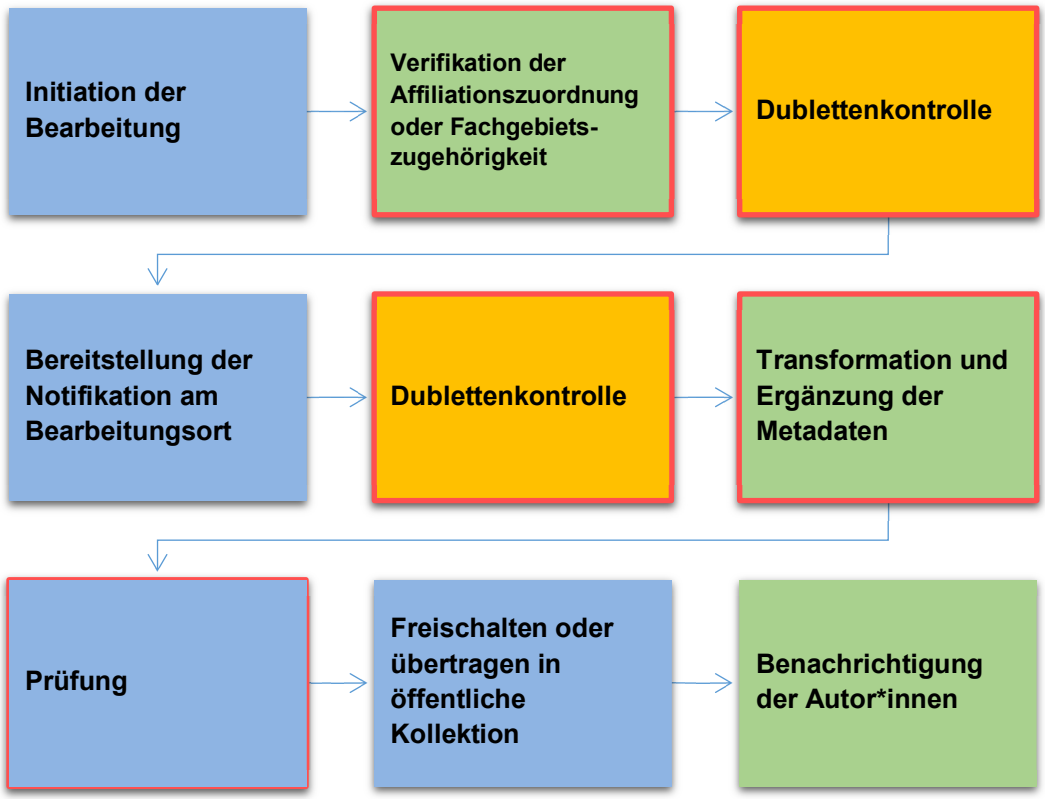

Abbildung 1: Abgeleiteter Workflow zur Bearbeitung von DeepGreen-Notifikationen

Arbeitsschritte in dieser Abbildung, die blau dargestellt sind, waren in allen evaluierten Workflows vorhanden. Grün dargestellte Arbeitsschritte wurden nicht in allen evaluierten Workflows eingesetzt. Einen Spezialfall stellt der Arbeitsschritt der Dublettenkontrolle dar. Einige Repositorien führen eine Dublettenkontrolle (orange) vor dem Import der DeepGreen-Notifikationen in den Bearbeitungsort durch, während andere erst die importierten

Publikationen auf Dubletten überprüfen. Die lizenzrechtliche Prüfung (roter Rahmen) wird von den verschiedenen Open-Access-Repositorien zu unterschiedlichen Zeitpunkten des Arbeitsablaufs durchgeführt.

Die Implementation der einzelnen identifizierten Arbeitsschritte wird von den an DeepGreen teilnehmenden Open-Access-Repositorien unterschiedlich umgesetzt. Die folgende Tabelle skizziert verschiedene Umsetzungsarten.

	Initiation der Bearbeitung • E-Mail an Bearbeiter*in • Eingang im Dashboard (regelmäßiges Einsehen) • Erinnerung im Kalender • Automatisch generierte Bearbeitungs-Tickets in einem vorhandenen Ticketing-System
Prüfung	**Verifikation der Affiliationszuordnung oder Fachgebietszugehörigkeit** • Manuelle Überprüfung
	Dublettenkontrolle • Verschiedene Workflows über den manuellen oder automatisierten Vergleich von Persistenten Identifikatoren (PIDs) oder Ähnlichkeit von Metadatenfeldern • Automatische Löschung oder Markierung zur manuellen Prüfung • Vor oder nach Import der Notifikationen in die Bearbeitungskollektion
	Bereitstellung der Notifikation am Bearbeitungsort • Automatisierter Import der Notifikationen von DeepGreen in eine Bearbeitungskollektion im Repositorium
Lizenzrechtliche	**Transformation und Ergänzung der Metadaten** • Eventuelle Umwandlung des Metadatenschemas (z.B. JATS -> MODS) • Ergänzung der Metadaten aus anderen Quellen wie z.B. Crossref via DOI von DeepGreen (automatisiert oder manuell) • Ergänzung durch eigene Metadaten (z.B. Import-Quelle "DeepGreen", zusätzliche Lizenzinformationen)
	Prüfung • Manuelle Prüfung der Affiliation • Manuelle Prüfung der automatisch identifizierten Dubletten • Manuelle Prüfung der Metadaten und der Zuordnung der Volltexte • Manuelles Anhängen der Volltexte • Überprüfung der Listung der Zeitschrift in DOAJ oder Policy-Check auf Sherpa Romeo
	Freischalten oder übertragen in öffentliche Kollektion • Freigabe durch weiteren Mitarbeiter (4-Augen-Prinzip) • verschieben in öffentliche Kollektion • ggf. Embargo einstellen
	Benachrichtigung der Autor*innen • E-Mail-Benachrichtigung

Tabelle 2: Implementationen der identifizierten Bearbeitungsschritte

4.1.1 Weitgehend automatisierter Workflow vs. manueller Workflow

Aus den evaluierten Workflows lassen sich zwei unterschiedliche Workflows ableiten, die sich hauptsächlich durch ihren Automatisierungsgrad unterscheiden. Weitgehend automatisierte Arbeitsabläufe laufen bis zum Punkt „Prüfung“ durch und es erfolgt lediglich eine manuelle Prüfung von Affiliation, Lizenzen und/oder Metadaten mit anschließender Freigabe.

Workflows, die hauptsächlich manuell erfolgen, fangen üblicherweise mit einer Prüfung der Affiliation oder einer lizenzrechtlichen Prüfung an, danach schließt sich die Dublettenkontrolle an. Dies erspart ein aufwändiges manuelles Ergänzen der Metadaten für irrelevante Notifikationen.

4.1.2 Besonderheiten des Workflows von Fachrepositorien

Aufgrund ihrer besonderen rechtlichen Situation und der spezifischen inhaltlichen Anforderungen ergeben sich für Open-Access-Fachrepositorien Abweichungen vom bisher diskutieren Arbeitsablauf. Im Gegensatz zu institutionellen Open-Access-Repositorien und Forschungsinformationssystemen sammeln Fachrepositorien nicht die Publikationen einer bestimmten Forschungseinrichtung, sondern bilden den Forschungsstand eines Fachgebiets ab.

Für Fachrepositorien entfällt daher der Arbeitsschritt *„Verifikation der Affiliationszuordnung“*. Die auf Basis der ISSN-Liste (siehe Unterkapitel *3.1.1 Zusatzanforderungen im Zuordnungsverfahren für Fachrepositorien*) erfolgte Zuordnung von Publikationen kann aber bei Bedarf durch eine detailliertere „*Verifikation der Fachgebietszugehörigkeit*“ ergänzt werden.

Dies kann entweder über eine manuelle Prüfung und Sichtung des Volltextes der Publikation erfolgen oder durch eine automatisierte Auswertung der Artikel-Metadaten. Um die automatische Verarbeitung der Notifikationen durch Fachrepositorien zu erleichtern stellt DeepGreen über das Web-API im Feld *„metadata.subject“* der Notifikation vom Verlag mitgelieferte Schlüsselwörter der Publikation aus den NLM/NISO JATS Feldern `<subject>` und `<kwd>` zur Verfügung.

Neben der fachlichen Zuordnung der Artikel müssen Open-Access-Fachrepositorien bei der Verarbeitung von Publikationen aus Allianz-, Nationallizenz oder FID-Lizenzen die lizenzrechtlichen Bedingungen beachten. Im Gegensatz zu institutionellen Open-Access-Repositorien und Forschungsinformationssystemen sind Fachrepositorien, die nicht zur Einrichtung der publizierenden Person gehören, nicht berechtigt das Open-Access-Recht in den Allianz-, Nationallizenz oder FID-Lizenzen im Namen der Autorin bzw. des Autors wahrzunehmen. Somit müssen sie den zusätzlichen Arbeitsschritt *„Zustimmung der Autor*innen einholen“* (Abbildung 2) berücksichtigen und – wenn sie nicht mit dem Verlag eine Kooperation eingehen – die Zustimmung zu der öffentlichen Zugänglichmachung im Fachrepositorium einzuholen.

4.2 Empfehlungen zum Workflow für Repositorien

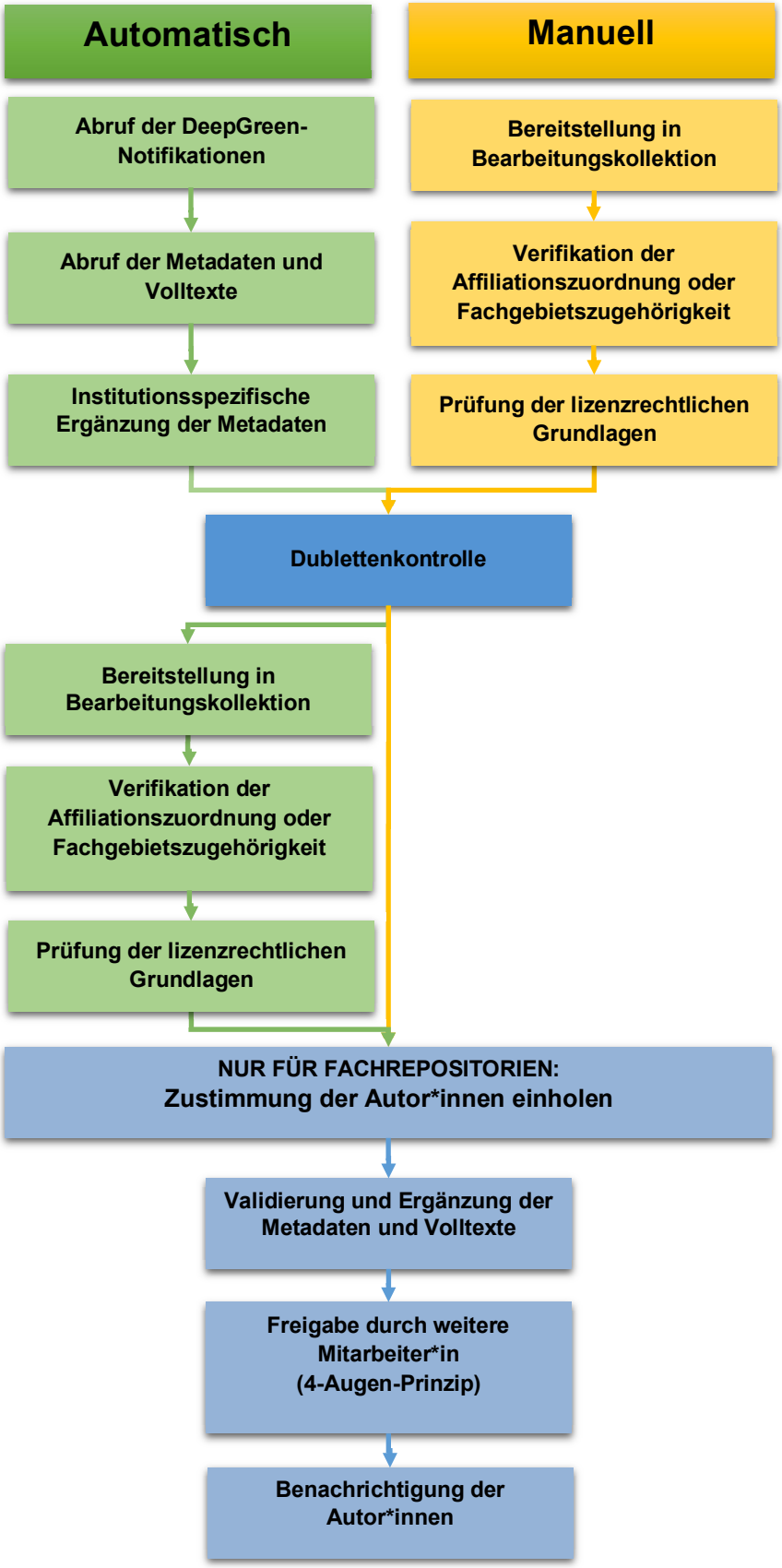

Abbildung 2: Schematische Darstellung empfohlener Workflows für die Verarbeitung von DeepGreen- Notifikationen durch Open-Access-Repositorien. Es werden die Grenzfälle einer weitgehend automatischen Verarbeitung und einer manuellen Verarbeitung sowie die zusätzlichen Arbeitsschritte für Fachrepositorien berücksichtigt.

Basierend auf den im vorherigen Abschnitt identifizierten gängigen Arbeitsschritten und den diskutierten Anwendungsfällen wurde ein allgemein zu empfehlender Workflow für Repositorien zur Verarbeitung von DeepGreen-Notifikationen abgeleitet. Der Workflow unterscheidet die beiden Verfahren einer vorwiegend automatisierten Verarbeitung und der manuellen Verarbeitung von Artikeln. Er berücksichtigt dabei die zusätzlichen Arbeitsschritte für Open-Access-Fachrepositorien.

Der Workflow mit vorwiegend automatisierter und der Workflow mit vorwiegend manueller Verarbeitung von Notifikationen unterscheidet sich in der Reihenfolge der Arbeitsschritte sowie im Arbeitsaufwand für den Arbeitsschritt *„Validierung und Ergänzung der Metadaten und Volltexte“*. Im manuellen Workflow liegt hier die Hauptlast in der manuellen Ergänzung der Metadaten. DeepGreen empfiehlt daher, den Workflow soweit wie möglich zu automatisieren. Die Dublettenkontrolle wurde im automatisierten Workflow nach dem automatischen Abruf der Metadaten und Volltexte positioniert, weil dadurch eine Nutzung der Metadaten und Volltexte – wie in Unterkapitel *4.3.2.3 Vergleich von Prüfsummen* diskutiert – ermöglicht wird.

4.3 Erkennung von Dubletten und anderer Dokumentversionen

4.3.1 Was sind Dubletten?

Naumann und Herschel (2010, S. 1) beschreiben Dubletten als „Multiple, yet different representations of the same real-world objects […]”. Aus Sicht von DeepGreen ist das „real-world object“ der zu einer Publikation von Verlagen gelieferte Datensatz, bestehend aus Metadaten und Volltext. Dieser wird von DeepGreen in Form einer Notifikation an die Open-Access-Repositorien weitergereicht.

Verschiedene Repräsentationen dieses Datensatzes können sich hinsichtlich der gelieferten Metadaten, aber auch hinsichtlich der gelieferten digitalen Objekte unterscheiden. So kann der Volltext in unterschiedlichen Dateiformaten durch DeepGreen zugestellt werden.
Ob ein Datensatz im Kontext einer betreffenden Institution als Dublette identifiziert werden soll oder nicht, hängt von den für das teilnehmende Open-Access-Repositorium definierten Anforderungen und Richtlinien ab. Um dies zu spezifizieren sind üblicherweise folgende Fragen zu beantworten:

- Auf welcher Strukturebene (Metadaten oder Volltext) sollen Dubletten identifiziert werden?

- Kann ein Volltext in unterschiedlichen Dateiformaten öffentlich zugänglich gemacht werden?

- Welche persistenten Identifikatoren werden von dem teilnehmenden Open-Access-Repositorium unterstützt, z.B. Digital Object Identifiers (DOIs), Uniform Resource Names (URNs)?

- Werden verschiedene Versionen derselben Publikation als eigenständige Datensätze betrachtet oder als ein Datensatz definiert?

- Wie werden verschiedene Sprachversionen derselben Publikation gehandhabt? Werden sie demselben Datensatz zugeordnet oder werden separate Datensätze erstellt?

- Sollen im Open-Access-Repositorium vorhandene Datensätze durch zusätzliche Informationen aus den Notifikationen angereichert werden?

    - Beispiel: Ein Datensatz im Repositorium enthält ausschließlich Metadaten und kann via DeepGreen mit einem Volltext aus der Notifikation aktualisiert werden.

Die Beantwortung dieser Fragen führt zu einer Entscheidung darüber, welche Parameter der von DeepGreen gelieferten Notifikationen mit den im Open-Access-Repositorium vorhandenen Datensätzen verglichen werden sollten, um potentiell im Repositorium gespeicherte Publikationen des von DeepGreen in der Notifikation gelieferten Datensatzes zu erkennen. Gängige Parameter zur Erkennung von Dubletten sind:

- Persistente Identifikatoren (DOI, URN) der Publikation

- Titel der Publikation

- Autor*innen der Publikation

- Prüfsummen über im Datensatz enthaltene digitale Objekte (z.B. Artikel-PDF), Metadaten oder den gesamten Datensatz

Die Kriterien, wann eine durch DeepGreen zugestellte Publikation als Dublette oder als eigenständige Publikation gewertet wird, muss jede Institution selbst definieren. In der Praxis gängige Kriterien sind:

- Identität aller oder einiger der überprüften Parameter (z.B. bei persistenten Identifikatoren)

- Ähnlichkeit der überprüften Parameter (z.B. bei Titel oder Autor*innen)

- Unterschied in einem bestimmten überprüften Parameter (z.B. Sprache einer Publikation) bei Identität anderer Parameter (z.B. persistente Identifikatoren)

- Überschreitung einer Vergleichsschwelle bei Vergleich mehrerer gewichteter Parameter

In der Praxis muss ein an DeepGreen teilnehmendes Open-Access-Repositorium für sich entscheiden, wie bei Erkennung einer Dublette vorgegangen werden soll.

4.3.2 Vorgehensweisen und technische Verfahren zur Dublettenerkennung

4.3.2.1 Vergleich von persistenten Identifikatoren

Die einfachste und schnellste Methode, um Dubletten zu identifizieren, besteht in einem Vergleich der persistenten Identifikatoren wie z.B. DOIs oder URNs. Diese Methode erfordert den Abgleich anhand der Identifikatoren mit dem eigenen Repositorium und kann daher weitgehend automatisiert werden.

4.3.2.2 Vergleich nicht eindeutiger Metadatenfelder

Metadatenfelder sind nicht eindeutig, weshalb ein Vergleich anhand dieser wenig hilfreich ist. In diesen Fällen empfiehlt sich eine vorherige Normalisierung der Metadatenfelder, siehe dazu die Veröffentlichung von Reichart und Mönnich (Reichart & Mönnich,1994, S. 198).

Eine gängige Praxis der Normalisierung ist die Umwandlung von Umlauten, Sonderzeichen oder zeichensatzspezifischen Zeichen, welche auf den direkten Vergleich wirken. Nach erfolgreicher Normalisierung kann grundsätzlich eine einfache Suche im Repositorium genutzt werden, um Dubletten zu erkennen. Dies setzt voraus, dass im Open-Access-Repositorium für jede Publikation automatisch auch die normalisierten Metadaten indiziert werden.

Bessere Ergebnisse beim Erkennen von Dubletten wurden durch die Verwendung von spezifischen Ähnlichkeitsfunktionen erzielt, welche fehlertolerant gegenüber Tippfehlern und Namensvariationen sind (Borges et al., 2011; Jiang et al., 2014) sowie durch die Verwendung von Machine-Learning-Algorithmen (Bilenko & Mooney, 2002; Bilenko & Mooney, 2003).

4.3.2.3 Vergleich von Prüfsummen

Ein weit verbreiteter Ansatz zur Erkennung von Dubletten ist die Verwendung von Prüfsummen (wie z.B. Checksums, Fingerprints, Hash Values, Signatures), insbesondere in Kombination mit deren Indizierung in integrierten Suchmaschinen. Prüfsummen können für Metadatenfelder oder deren Kombinationen, digitale Objekte oder auch für Publikationen erstellt werden.

Die am häufigsten genutzten Suchmaschinen für Repositorienplattformen Apache Solr (Lucene, 2020a) der Apache Software Foundation und Elasticsearch (Elastic, 2020a) des Elasticsearch BV bieten unterschiedliche Unterstützung für einen Dublettenerkennungsworkflow unter Verwendung von Prüfsummen an. Apache Solr stellt bereits spezialisierte Klassen (Lucene, 2020b) zur Verfügung, um diesen Workflow zu unterstützen. Für Elasticsearch kann entweder das zusätzliche, frei verfügbare Tool Logstash oder das Painless Scripting API (Elastic, 2020b) genutzt werden, um Prüfsummen (Elastic, 2020c) automatisch erstellen und indizieren zu können.

Auf diesem Wege werden bei beiden Suchmaschinen bei der Registrierung eines neuen Datensatzes (oder einer neuen Notifikation) automatisch, gemäß einer individuellen Auswahl und / oder Kombination von Metadatenfeldern, Prüfsummen erstellt und indiziert, welche dann über eine einfache Suche zur Erkennung von Dubletten genutzt werden können.

Gängige Prüfsummen-Algorithmen, die in diesem Kontext Verwendung finden und von beiden Suchmaschinen unterstützt werden, sind der Message-Digest Algorithm 5 (MD5) und der Secure Hash Algorithm (SHA1). Apache Solr stellt weiterhin den fuzzy-hashing Algorithmus Lookup3 zur Verfügung, welcher insbesondere für längere Texte geeignet ist und zudem near duplicates findet, während mit Elasticsearch/Logstash eine Reihe weiterer Prüfsummen generiert werden können.

4.3.3 Dubletten in verschiedenen Software-Lösungen für Repositorien

4.3.3.1 DSpace

DSpace ist eine Open-Source-Software zum Betrieb eines Dokumentenservers, die ursprünglich von Hewlett-Packard (HP) und dem Massachusetts Institute of Technologies (MIT) gemeinsam entwickelt wurde. Heute wird DSpace durch die nicht-kommerziellen Organisationen Duraspace und Lyrasis weiterentwickelt (Lyrasis DuraSpace, 2020). Die im Februar 2021 aktuelle Version von DSpace ist 6.3. DSpace 7 befindet sich noch in der Beta-Phase. Keine der beiden Version enthält eine integrierte Dublettenerkennung. Es gibt jedoch Erweiterungen von DSpace, die verschiedene Formen der Dublettenerkennung zur Verfügung stellen oder unterstützen. Für DSpace 6.3 steht das Add-On „Duplicate Detection Service for DSpace 6 (JSPUI)“ der Firma The Library Code frei zur Verfügung, welches auf der Analyse der Ähnlichkeit von Titel-Metadaten basierend auf der Levenshtein Editierdistanz beruht (Lyrasis DuraSpace, 2019).

Eine grundlegende Dublettenerkennung ist auch mit der DSpace-Erweiterung DSpace-CRIS verfügbar. Diese basiert auf dem Vergleich von Prüfsummen, welche in Apache Solr indiziert werden und steht sowohl bei der Einreichung von neuen Datensätzen also auch im Administrationsinterface zur Verfügung. Die für die Generierung der Prüfsummen genutzten Metadatenfelder können individuell konfiguriert werden. Die Dublettenerkennung wird auch mit der Veröffentlichung von DSpace 7 eine Erweiterung bleiben und nicht in das DSpace-Kernmodul integriert werden. Die Weiterentwicklung wird aber im Rahmen von DSpace-CRIS 7 vorangetrieben. Da die in DSpace-CRIS 7 implementierten Funktionalitäten frei verfügbar sind, könnten die Funktionalitäten zur Dublettenerkennung von den Open-Access-Repositorien bei Bedarf eigenständig in DSpace 7 integriert werden.

4.3.3.2 MyCoRe

MyCoRe (My Content Repository) ist ein Open-Source-Framework zum Erstellen von Repository-Webanwendungen, welche individuell implementiert werden. MyCoRe ist aus dem Verbundprojekt MILESS der Universität Duisburg-Essen hervorgegangen und wird von einer Entwickler-Community der deutschen Universitäten weiterentwickelt (MyCoRe, 2020).

Das MyCoRe-Framework beinhaltet derzeit keine Funktionalitäten zur Dublettenerkennung. Da MyCoRe aber auch Apache Solr als Suchmaschine nutzt, empfiehlt sich auch hier der oben beschriebene Ansatz zur Dublettenerkennung durch die Indizierung und den Vergleich von Prüfsummen.

4.3.3.3 OPUS

OPUS ist eine Open-Source-Software für den Betrieb von institutionellen und fachlichen Repositorien, welche vom Kooperativen Bibliotheksverbund Berlin-Brandenburg (KOBV) mit Unterstützung des Bibliotheksverbund Bayern (BVB) sowie des Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) weiterentwickelt wird (KOBV, 2020).

Die derzeit aktuelle Version OPUS 4 bietet zwar Funktionen zur Erstellung von Prüfsummen über digitale Objekte die zur Überprüfung der Dokumentenintegrität genutzt werden, aber keine dezidierte Dublettenerkennung. Auch OPUS 4 verwendet Apache Solr als integrierte Suchmaschine, deren Funktionalitäten entsprechend für die Dublettenerkennung genutzt werden können.

4.3.3.4 LibreCat

LibreCat ist eine Repositorien-Software, die durch die Universitäten Bielefeld, Gent und Lund weiterentwickelt wird. LibreCat verfügt mit der Version 2.0.3 über eine einfache Dublettenerkennung auf Basis von Persistenten Identifikatoren. Da LibreCat außerdem Elasticsearch (Elastic, 2020a) als integrierte Suchmaschine verwendet, stehen die oben diskutierten Möglichkeiten zur Dublettenerkennung mittels Prüfsummen auch hier zur Verfügung (LibreCat & Catmandu, 2020).

4.3.3.5 EPrints

EPrints ist eine Software für Repositorien, die durch die Universität Southampton entwickelt wurde (EPrints, 2020a). EPrints bietet zwar keine dezidierte Dublettenerkennung, das integrierte *Issue Detection System* kann aber durch verfügbare Plugins zur Erkennung von identischen oder ähnlichen Titeln erweitert werden. EPrints verwendet in der aktuellen Version 3.4.1 (EPrints, 2020b) eine proprietäre Suchmaschine, welche perspektivisch in einer geplanten neuen Version durch Xapian oder Apache Solr ersetzt werden könnte.

4.3.3.6 Zusammenfassung

Abschließend kann festgestellt werden, dass die hier betrachteten Software-Lösungen bisher keine einheitlichen Mechanismen zur Dublettenerkennung implementiert haben. Eine Übersicht über die integrierten oder durch Erweiterungen zur Verfügung gestellten Funktionen sind in der untenstehenden Tabelle 3 zu finden. Da bis auf EPrints alle Software-Lösungen Suchmaschinen verwenden, welche Mechanismen zur Erstellung von Prüfsummen zur Verfügung stellen, bietet sich in diesen Fällen die eigene Implementierung einer Dublettenerkennung auf Basis von Prüfsummen an. Grundlegende Workflows dafür werden im folgenden Abschnitt skizziert.

Plattform	Version	Integrierte Suchmaschine	Integrierte Mechanismen der Dublettenerkennung
DSpace	6.3	Apache Solr	AddOn zum Erkennen von ähnlichen Titeln
	7 (beta)		
DSpace - CRIS	5.10	Apache Solr	
	6.3		AddOn zum Erkennen von ähnlichen Titeln
	7 (beta)		Vergleich von Prüfsummen
EPrints	3.4.1	proprietär	Plugins zum Erkennen von doppelten oder ähnlichen Titeln mittels „Issue Detection System“
LibreCat	2.0.3	Elasticsearch	Dublettenerkennung auf Basis von Persistenten Identifikatoren (DOI, ISI, PMID, arxiv)
MyCoRe	2019.06 (LTS)	Apache Solr	
Opus	4.6	Apache Solr	

Tabelle 3: Übersicht Dublettenerkennung in verschiedenen Software-Lösungen für Repositorien

4.3.4 Allgemeine Workflows zur Dublettenerkennung

Aus den oben zusammengefassten Funktionen der einzelnen Software-Lösungen für Repositorien können, je nach integrierter Suchmaschine, verschiedene allgemeine Workflows zur Dublettenerkennung entwickelt werden. Grundsätzlich empfiehlt sich ein Workflow über die Generierung und Indizierung von Prüfsummen, da die in den meisten Software-Lösungen integrierten Suchmaschinen Funktionen bieten, die diesen Workflow bereits unterstützen. Die folgenden drei Abbildungen skizzieren empfohlene Workflows für die Suchmaschinen Apache Solr, Elasticsearch und Elasticsearch in Kombination mit Logstash.

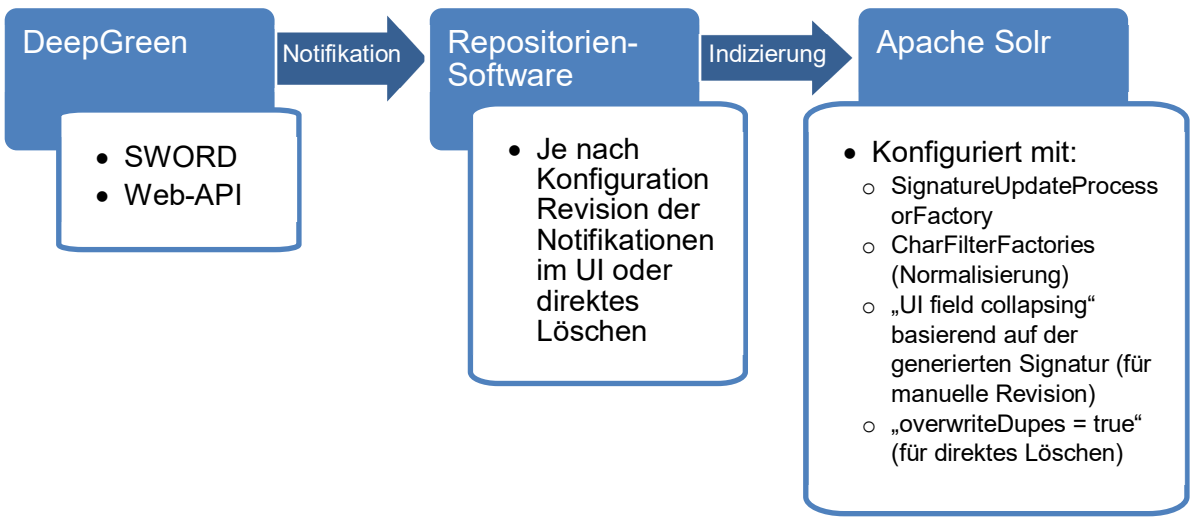

Abbildung 3: Workflow zur Dublettenerkennung mit Apache Solr

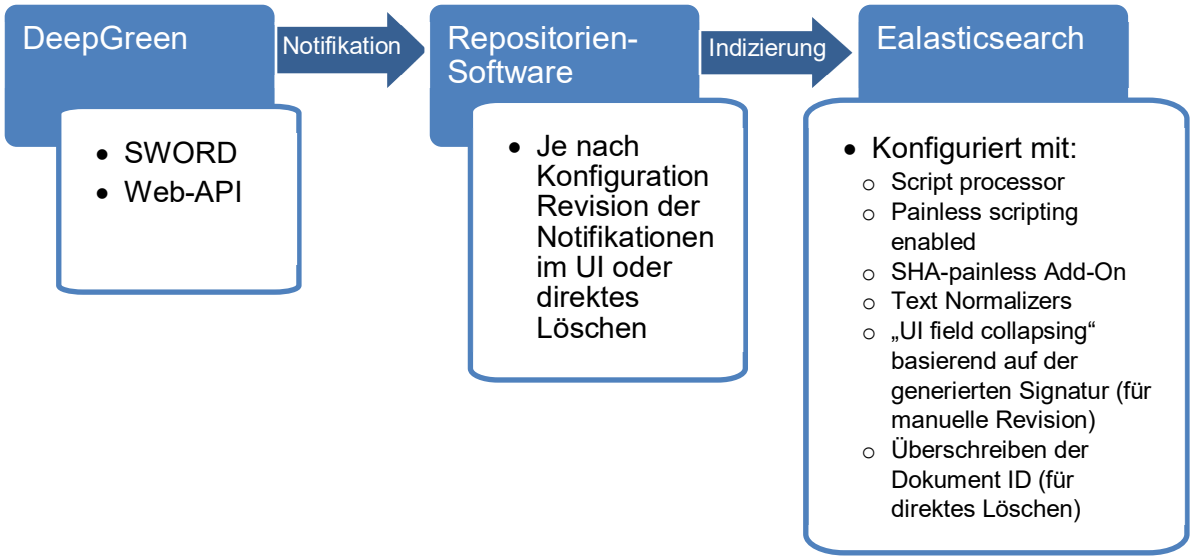

Abbildung 4: Workflow zur Dublettenerkennung mit Elasticsearch

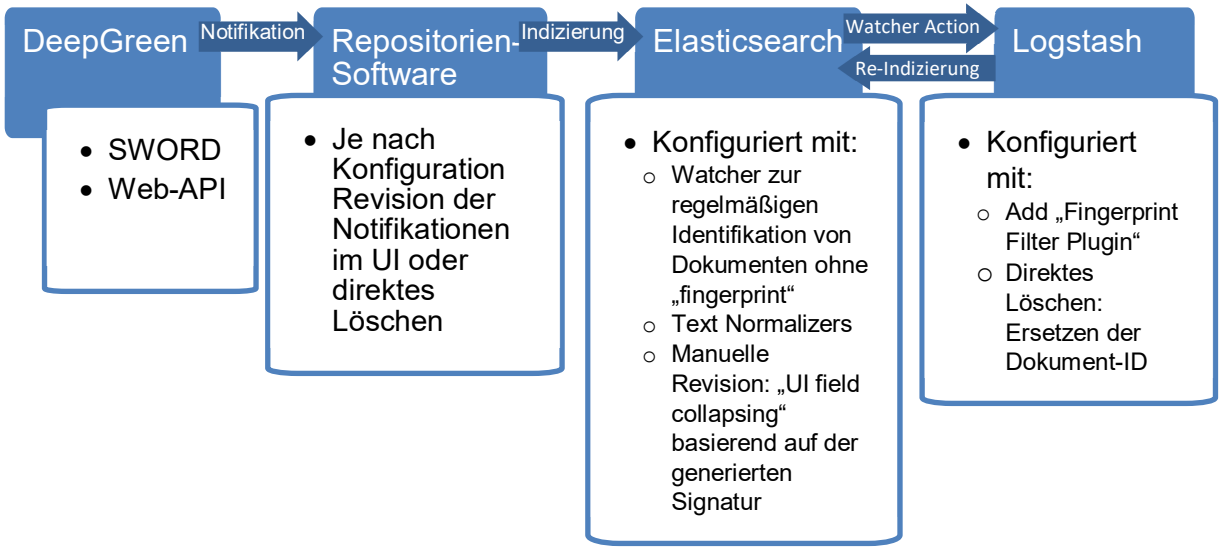

Abbildung 5: Workflow zur Dublettenerkennung mit Elasticsearch und Logstash

4.4 Umgang mit Embargofristen

DeepGreen bezieht Embargoinformationen zu Notifikationen für Allianz- und Nationallizenzen in Form von Monaten über die EZB. Auch bei der Verteilung nach Verlagspolicy können DeepGreen-Notifikationen prinzipiell Embargoinformationen enthalten, falls diese von den Verlagen zur Verfügung gestellt werden.
Da es derzeit keine etablierten Standards für die Einbettung von Embargoinformationen in Metadatenformate gibt, stellt DeepGreen Embargoinformationen ausschließlich über das Web-API im Feld *„embargo.duration“* zur Verfügung. Die Metadaten einer Notifikation können für jede Notifikation bei Angabe der `<Notification-ID>` über die URL

`https://www.oa-deepgreen.de/api/v1/notification/<Notification-ID>`

abgerufen werden. Der Parameter `<Notification-ID>` muss dabei durch die entsprechende Kennung (ID) der Notifikation ersetzt werden (vgl. Unterkapitel *4.3.1 Was sind Dubletten?*).

Es liegt in der Verantwortung der Open-Access-Repositorien, die Richtigkeit der Angaben zu überprüfen sowie das exakte Zweitveröffentlichungsdatum zu bestimmen.

5 Perspektiven zu potenziellen Weiterentwicklungen für den Dienst DeepGreen

5.1 Zuordnung zu Fachrepositorien

Die fachliche Zuordnung von Artikeln zu Fachrepositorien erfolgt mit Stand Februar 2021 anhand von individualisierten ISSN-Listen. Im Rahmen des Projekts DeepGreen wurden jedoch auch andere Vorgehensweisen evaluiert, die eventuell in zukünftigen Weiterentwicklungen realisiert werden können.

Einen Ansatzpunkt zur Fachzuordnung bieten die Keywords in NISO JATS. Allerdings sind die verschiedenen Elemente zur Auszeichnung von Keywords in der NLM JATS DTD weder obligatorisch noch inhaltlich standardisiert. Das heißt, dass das XML auch ohne Keywords valide ist und sowohl in den internen Verlagssystemen als auch über Schnittstellen prozessiert werden kann. So können auch Verlage an DeepGreen teilnehmen, die keine Keywords erfassen oder eben Artikel in DeepGreen prozessiert werden, in denen die Autor*innen keine Keywords beigetragen haben. Die fehlende Standardisierung erhöht zudem die Fehleranfälligkeit von Matchingprozessen und die Wahrscheinlichkeit, dass Keywords nicht zugeordnet werden können. Diese Felder stellen daher keinen idealen Ausgangspunkt für eine fachspezifische Zuordnung von Artikeln dar.

5.2 Anforderungen der Repositorien an Verlagsdatenlieferungen

Ein wesentlicher Mehrwert von DeepGreen für Open-Access-Repositorien besteht in einer Reduktion des Arbeitsaufwands durch die gezielte Zuordnung von relevanten Publikationen. Dieser Mehrwert wird durch die Affiliationszuordnung von DeepGreen gewährleistet, welche jedoch aufgrund der vielen möglichen Namensvariationen einer Institution Herausforderungen mit sich bringt. Bessere Zuordnungsverfahren basierend auf standardisierten Daten, z.B. bezüglich institutioneller Affiliationen, wären hier sehr hilfreich.

In einer Umfrage von DeepGreen unter wissenschaftlichen Verlagen zeigten sich aber bereits die oben genannten Hürden in der einwandfreien Zuordnung der Artikel anhand der Affiliationsangaben als Textstrings.

Eine Empfehlung von DeepGreen an wissenschaftliche Institutionen und Verlage ist es daher, den Grad der Standardisierung zu erhöhen und diesem Themenfeld im Dialog mit Verlagen mehr Aufmerksamkeit zu schenken. Der Fokus sollte hier auf einer aktiven Förderung der Nutzung von Persistent Identifiers liegen. Metadatenfelder für Identifier sind in der NISO JATS DTD bereits vorgesehen: Die Contributer ID durch das Element `<contrib-id-type>` und die Institution ID durch `<institution-id>`. Diese sollten möglichst in der Eingabemaske des Submission Systems abgefragt und vor allem auch sogleich automatisch zugeordnet werden, sodass keine Fehler in den Affiliationsangaben der Publizierenden vorkommen können. Da außerdem verschiedene Abrechnungsmodelle, Lizenzen, Vereinbarungen oder

Verträge zwischen Verlagen und Bibliotheken Einfluss haben, sollte die eindeutige Zuordenbarkeit der Artikel im beidseitigen Interesse sein.

DeepGreen plädiert für eine engere Zusammenarbeit auf diesem Gebiet und intensiveren Austausch zwischen allen Akteuren über Bestrebungen, Projekte und Lösungsvorschläge in der Branche. Ein gutes Verständnis über die Arbeitsweise der jeweils anderen Akteure ist bei Projekten wie DeepGreen elementar. Alle Akteure sollten gemeinsam über das Problem der eindeutigen fachlichen Zuordenbarkeit beraten. Einen Ansatz bietet die Listung von Anforderungen an Verlagsmetadaten (Pampel 2019, S. 9-10). Auch wissenschaftliche Einrichtungen sollten sicherstellen, dass die Qualität der Metadaten deutlich verbessert wird. Eine standardisierte und gemeinsame Vorgehensweise in diesem Bereich wäre ein wichtiger Vorstoß, von dem die gesamte wissenschaftliche Publikationslandschaft profitieren würde.

Literaturhinweise

Letztes Abrufdatum der Internet-Dokumente innerhalb der Fußnoten und Literaturhinweise: 26.01.2021

Bilenko, M., & Mooney, R. J. (2002): „Learning to combine trained distance metrics for duplicate detection in databases“ (Technical Report AI 02-296). Artificial Intelligence Laboratory, University of Texas at Austin, Austin TX

Bilenko, M., & Mooney, R. J. (2003): „Adaptive duplicate detection using learnable string similarity measures“. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*: S. 39-48. KDD ’03. Washington, D.C.: Association for Computing Machinery. https://doi.org/10.1145/956750.956759

Borges, E. N., de Carvalho, M. G., Galante, R., Gonçalves, M. A., & Laender, A. H. F. (2011): „An Unsupervised Heuristic-Based Approach for Bibliographic Metadata Deduplication“. Information Processing & Management, Managing and Mining Multilingual Documents, 47 (5): S. 706-718. https://doi.org/10.1016/j.ipm.2011.01.009

DeepGreen (2020a): Projektkommunikation. https://deepgreen.kobv.de/de/deepgreen/pr/

DeepGreen (2020b): Anleitung zur Benutzung von DeepGreen. https://oa-deepgreen.github.io/user_docs/index.html

DFG (2014): Open-Access-Transformation. https://www.dfg.de/foerderung/info_wissenschaft/2014/info_wissenschaft_14_29/index.html

DFG (2015): Grundsätze für den Erwerb DFG-geförderter überregionaler Lizenzen (Allianz-Lizenzen), S. 9. http://www.dfg.de/formulare/12_181/12_181_de.pdf

Elastic (2020a): Elasticsearch. https://www.elastic.co/de/elasticsearch/

Elastic (2020b): Painless API Reference. https://www.elastic.co/guide/en/elasticsearch/painless/master/painless-api-reference.html

Elastic (2020c): Fingerprint filter plugin. https://www.elastic.co/guide/en/logstash/current/plugins-filters-fingerprint.html

EPrints (2020a): About EPrints. https://www.eprints.org/uk/index.php/about/

EPrints (2020b): What’s new in 3.4.1? https://www.eprints.org/uk/index.php/whats-new-in-3-4-1/

EZB (2016): OA-EZB-Schnittstelle. http://ezb.uni-regensburg.de/ezeit/services/oa-ezb.phtml?bibid=UBR&colors=7&lang=de https://www.nationallizenzen.de/open-access/open-access-rechte.xls/view

Jiang, Y., Lin, C., Meng, W., Yu, C., Cohen, A. M., & Smalheiser, N. R. (2014): „Rule-Based Deduplication of Article Records from Bibliographic Databases“. *Database* 2014 (Januar). https://doi.org/10.1093/database/bat086

KOBV (2020): OPUS4. https://www.kobv.de/entwicklung/software/opus-4/

LibreCat & Catmandu (2020): Open source applications for libraries. https://librecat.org/

Lucene (2020a): Solr. https://lucene.apache.org/solr

Lucene (2020b): De-Duplication. https://lucene.apache.org/solr/guide/8_5/de-duplication.html

Lyrasis DuraSpace (2019): Enhance your DSpace Installation with Free Plugins from The Library Code. https://duraspace.org/enhance-your-dspace-installation-with-free-plugins-from-the-library-code/

Lyrasis DuraSpace (2020): About DuraSpace / History. https://duraspace.org/about/history/

MyCoRe (2020): Was kann MyCoRe? https://www.mycore.de/site/features/

Naumann, F., & Herschel, M. (2010): „An Introduction to Duplicate Detection“. *Synthesis Lectures on Data Management* 2 (1): S. 1-87. https://doi.org/10.2200/S00262ED1V01Y201003DTM003

Pampel, H. (2019): Auf dem Weg zum Informationsbudget: zur Notwendigkeit von Monitoringverfahren für wissenschaftliche Publikationen und deren Kosten; Arbeitspapier, Potsdam : Helmholtz Open Science Koordinationsbüro, S. 9-10. https://doi.org/10.2312/os.helmholtz.006

Putnings, M., & Rusch, B. (2016): DeepGreen - Entwicklung eines rechtssicheren Workflows zur effizienten Umsetzung der Open-Access-Komponente in den Allianz-Lizenzen für die Wissenschaft. *o-bib. Das offene Bibliotheksjournal / herausgegeben vom VDB*, Bd. 3: S. 110-119. https://doi.org/10.5282/O-BIB/2016H4S110-119

Reichart, M., & Mönnich, M. W. (1994): „Dublettenkontrolle in bibliographischen Datenbanken“. *Bibliothek Forschung und Praxis* 18 (2): S. 193-216. https://doi.org/10.1515/bfup.1994.18.2.193

Stöber, A. (2012): Open-Access-Rechte in Allianz- und Nationallizenzen: Eine Handreichung für Repository-Manager, Bibliothekare und Autoren, (Arbeitsgruppen „Nationale Lizenzierung“ und „Open Access“ der Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen, Ed.), 17 S. https://doi.org/10.2312/allianzoa.004

Verbundzentrale des GBV (2019): Exceltabelle Übersicht zur Nutzung der verhandelten Open-Access-Rechte. https://www.nationallizenzen.de/open-access/open-access-rechte.xls/view