Konzeptpapier ORCID DE Monitor

Friedrich Summann, Universitätsbibliothek Bielefeld
Stephanie Glagla-Dietz, Deutsche Nationalbibliothek Frankfurt am Main
Sebastian Wolf, Universitätsbibliothek Bielefeld

Im Rahmen des DFG-Projektes **ORCID DE 2 – Konsolidierung der ORCID-Informationsinfrastruktur in Deutschland**[1] ist ein Arbeitspaket beantragt und bewilligt worden, das eine Monitorfunktion zur Beobachtung und Auswertung der Entwicklung der ORCID-Nutzung bereitstellen soll. Damit soll es insbesondere ermöglicht werden, die Entwicklung der im Rahmen des Projektes verfolgten Ziele zu überprüfen und auch die Ableitung von Strategien zur Feinjustierung der Projektmaßnahmen zu unterstützen.

Der ORCID DE Monitor soll umfassend die Verwendung der ORCID iD in unterschiedlichsten Publikationsumgebungen dokumentieren und deren zunehmende Verbreitung nachweisen.

Insbesondere sind die eigens in der ersten Phase des Projektes[2] entwickelten Analyse-Skripte, die die globale (per Stichproben) und nationale (komplette) Repositorienlandschaft evaluieren, zu integrieren. Diese Skripte werden im Hinblick auf einen periodisch automatisierten Ablauf (je nach Laufzeit z. B. monatlich oder vierteljährlich) optimiert. Gleichzeitig wird die Datenstruktur fortentwickelt, um die verschiedenen heterogen vorliegenden Daten zu normalisieren und die Abfragemöglichkeiten optimal zu gestalten. Die bereits vorhandenen Daten (erhoben seit Start der ersten Phase des Projektes im Jahr 2016) sollen für die ORCID DE Monitor Datenbasis mitberücksichtigt werden. Damit und auf Grundlage der über teilweise umfangreiche Zeiträume vorliegenden Zahlen werden insbesondere Timeline-Auswertungen möglich, die Entwicklungen dokumentieren und aufzeigen können.

Aktuell stehen Daten aus den folgenden Datenquellen zur Verfügung:

- **Vorkommen in Repositorien (national / global) in Metadaten**

Die Daten werden dabei mit eigens entwickelten Skripten ermittelt und abgelegt. Seit Beginn des Projekts ORCID DE liegt ein Skript vor, das die BASE-Liste der globalen Repositorien verwendet und versucht via OAI-PMH durch Stichprobenaufrufe in verschiedenen Metadatenformaten die Verwendung von ORCID iDs aufzufinden und zu protokollieren. Bisher ist dieses Skript (die Laufzeit erstreckt sich über mehrere Tage) gemittelt zwei Mal pro Jahr eingesetzt worden und daher liegen Daten in unterschiedlichen Abständen seit Mai 2016 vor.

Insbesondere werden ausgegeben:

---

[1] siehe Bertelmann, R., Cruse, P., Niggemann, E., Pieper, D., Sens, I., Burger, M., Dasler, R., Dreyer, B., Elger, K., Fenner, M., Hagemann-Wilholt, S., Hartmann, S., Höhnow, T., Kett, J., Pampel, H., Pietsch, C., Schirrwagen, J., Summann, F. (2019): ORCID DE 2 – Konsolidierung der ORCID-Informationsinfrastruktur in Deutschland, 21 p. https://doi.org/10.2312/lis.20.01

[2] siehe Bertelmann, R., Niggemann, E., Pieper, D., Elger, K., Fenner, M., Hartmann, S., Höhnow, T., Jahn, N., Müller, U., Pampel, H., Schirrwagen, J., Summann, F. (2015): ORCID DE – Förderung der Open Researcher and Contributor ID in Deutschland, 24 p. https://doi.org/10.2312/lis.16.01

- ★ Gesamtzahl Publikationen mit ORCID iDs pro Datenquelle
- ★ Anzahl Publikationen mit ORCID iDs pro Plattformsystem
- ★ Anzahl Publikationen mit ORCID iDs pro Metadatenformat

Damit lassen sich Timelines für die allgemeine Verbreitung, bezogen auf die Merkmale technischer Plattformsysteme und verwendete Metadatenformate erstellen. Vorgenommen wurde auch eine länderbezogene Auswertung, die eine vergleichende Analyse mit anderen Repository-Netzwerken erlaubt.

- **Vorkommen in indexierten BASE-Metadaten**

Die Daten liegen als Facetten, bereitgestellt vom BASE-Index, vor und beinhalten neben der Gesamtzahl die Zahlen für Publikationen mit ORCID iDs pro BASE-Datenprovider. Attribute der Datenprovider können aus der sog. Admin-Datenbank (der Datenlieferanten) dazugespielt werden. Die Daten werden im Falle von Repositorien überwiegend aus den Publikationsmetadaten auf Basis des Dublin Core-Formats und im Falle Crossref Member auf Basis des Crossref-Metadatenformats im Rahmen der Datennormalisierung ermittelt und in einem dezidierten Suchfeld abgelegt.

Beispiel aus den Rohdaten (Stand 1.8.2020):


```
"f_dccollection": [
[       "crelsevierbv",
        521029          ],
[               "crwiley",
        516498          ]
[               "ftunivqespace",
        230393          ],
[               "crsagepubl",
        86811           ],
[               "crplos",
        71808           ],
[               "crf1000",
        67417           ],
[               "ftspringeroc",
        50593           ],
[               "crcambridgeupr",
        27810           ],
[               "ftawi",
        24403           ],
```


Die Attributinformationen zur Quelle (Name, Land, Repositorytyp, Systemplattform etc.), in der Datenstruktur durch den BASE-Collection-Identifier eindeutig zugeordnet, werden aus der BASE-Metadaten-DB hinzugefügt. Im obigen Beispiel steht z.B. crwiley für Wiley via Crossref und ftawi für das Repository des Alfred-Wegener-Instituts, Helmholtz-Zentrum für Polar- und Meeresforschung (AWI).

- **Vorkommen in BASE Claiming Service**

Es liegen die Zahlen vor für:

★ "all_users":
★ "claiming_users"
★ "claimed_docs":
★ "gnd_ids"

- **ORCID-Vorkommen in Crossref-Daten**

Die Crossref API liefert zu jedem Crossref-Mitglied (Crossref Member) Zahlen zur Gesamtanzahl der Metadaten als auch Prozentzahlen zum Anteil von ORCID iDs im laufenden Jahrgang wie auch in den Backfiles. Eine Länderzuordnung für jeden Member ist in den Adressdaten in normierter Form vorhanden. Ein interessanter Auswertungsaspekt dabei ist, eine Zuordnung nach Typ des jeweiligen Crossref Members, um insbesondere zwischen den gängigen Ausprägungen kommerzielle Verleger, Organisationen, akademischen Institutionen, Zeitschriftenhosts und Einzelzeitschriften unterscheiden zu können. Allerdings liegen diese Daten bei Crossref nicht vor und daher müssen diese Angaben im Rahmen des Projekts zusammengestellt und ergänzt werden. Wegen des damit verbundenen hohen Aufwandes für die Zuordnung, wird sich diese Ergänzung vermutlich auf Quellen aus der Bundesrepublik Deutschland beschränken müssen.

Aktuell werden (im BASE-Umfeld) für jeden Crossref Member unter anderem monatlich via API abgeholt und für spätere Zugriffe abgelegt

- Anzahl Objekte (Crossref-Angabe)
- Land
- Prozentzahl ORCID iDs in Objekten für aktuellen Jahrgang
- Prozentzahl ORCID iDs in Objekten für Backfiles (berechnet über Anzahl Objekte)

- **Vorkommen in der Gemeinsamen Normdatei (GND)**

Die DNB liefert monatlich Zahlen zu den GND-Personendatensätzen, die eine ORCID iD als Standardnummer in einem dafür vorgesehenen Feld enthalten. Sie werden seit Mai 2016 manuell eingetragen. Ab März 2019 wurden in unterschiedlichen Abgleichverfahren übereinstimmende Personendatensätze ermittelt und bisher insgesamt 63.547 ORCID iDs in die GND eingespielt (s. [6]). Die 100.000 iD-Schwelle wurde am 18.12.2020 überschritten (s. [7]).

- **Claiming-Service in der Deutschen Nationalbibliographie**

Seit Juli 2019 ist es allen ORCID-Nutzer*innen möglich, ihre in der Deutschen Nationalbibliografie gelisteten Publikationen aus ihren ORCID-Records heraus zu claimen (s. [6] und [8]). An den ORCID DE Monitor werden die Anzahl der Claimenden (ORCID-Nutzer*innen nach Bereinigung, also immer

nur der erste Claiming-Vorgang eines/r ORCID-Nutzers/in), die Anzahl der geclaimten GND-IDs (vor Bereinigung, also zum Claiming-Zeitpunkt) und die Anzahl der geclaimten Titeldatensätze der Deutschen Nationalbibliografie (nach Bereinigung, also ohne mehrfache Claiming-Vorgänge derselben Titel) geliefert.

- **Standardnummern in Titeldatensätzen der Deutschen Nationalbibliografie**

Hier werden Zahlen der Titeldaten im DNB-Katalog verwendet, die eine oder mehrere ORCID iDs oder ISNI IDs enthalten.

- **ORCID-Zahlen**

Via ORCID-Dashboard werden Angaben zur ORCID-Community, insbesondere zu den registrierten Wissenschaftler*innen ausgewertet. Es werden Transferskripte entwickelt, die die aktuell in heterogenen Datenformaten vorliegenden Basis-Daten der unterschiedlichen Quellen in ein einheitliches Format verwandeln und für den API-Zugriff in einem angekoppelten Datenspeicher abgelegt (s. die geplante Infrastruktur in Abb. 1).

Die folgende Tabelle zeigt in der Übersicht die verfügbaren Daten, ihre Herkunft und die zeitliche Verfügbarkeit.

Datentyp	**Datenquelle**	**Daten verfügbar seit**
Anzahl ORCID iDs in Repository-Metadaten	Evaluierungsskripte in OAI-PMH-Repositorymetadaten	2016
ORCID iDs in Crossref-Metadaten	Crossref API	2017
ORCID-Vorkommen in BASE Metadaten	BASE Index (Facetten)	2020 (Jan)
ORCID-Vorkommen in BASE Claiming	BASE Claiming Datenbank	2019
Anzahl Wissenschaftler*innen mit ORCID	ORCID	2016
GND-Datensätze mit ORCID iD	GND	2016 (Mai)
Claiming-Service in der Deutschen Nationalbibliographie	DNB	2019 (Juli)
ORCID iDs in Titeldaten der Deutschen Nationalbibliografie	DNB-Katalog	-

Zusätzlich liegt Zahlenmaterial länder-bezogen vor, das im BASE-Umfeld zur Beschreibung der Wissenschaftsinfrastruktur vorgehalten wird. Dazu gehören Angaben zur Anzahl der Wissenschaftler*innen, Budgetangaben für die Wissenschaft, Anzahl der Publikationen in einem Land. Diese Angaben sind länderbasierten OECD-Datenangaben entnommen und normalisiert worden (weil sie z. B. nicht durchgehend für denselben Zeitraum vorliegen) und ermöglichen es insbesondere, Daten für einen vergleichenden Ansatz zu liefern und damit in Abhängigkeit von den genannten Größen für den Ländervergleich zu relativieren und somit vergleichbarer zu machen.

All diese Daten werden basierend auf der an der UB Bielefeld existierenden Daten-Infrastruktur aus dem BASE-Kontext (s. Abb. [1]) integriert und bereitgestellt. Zusätzlich wird mit Hilfe der erweiterten API eine eigene projektbezogene Visualisierungsinstanz des ORCID DE Monitors (Client 1 in Abb. 1) realisiert. Die komplette Infrastruktur zeigt die Abbildung 1.

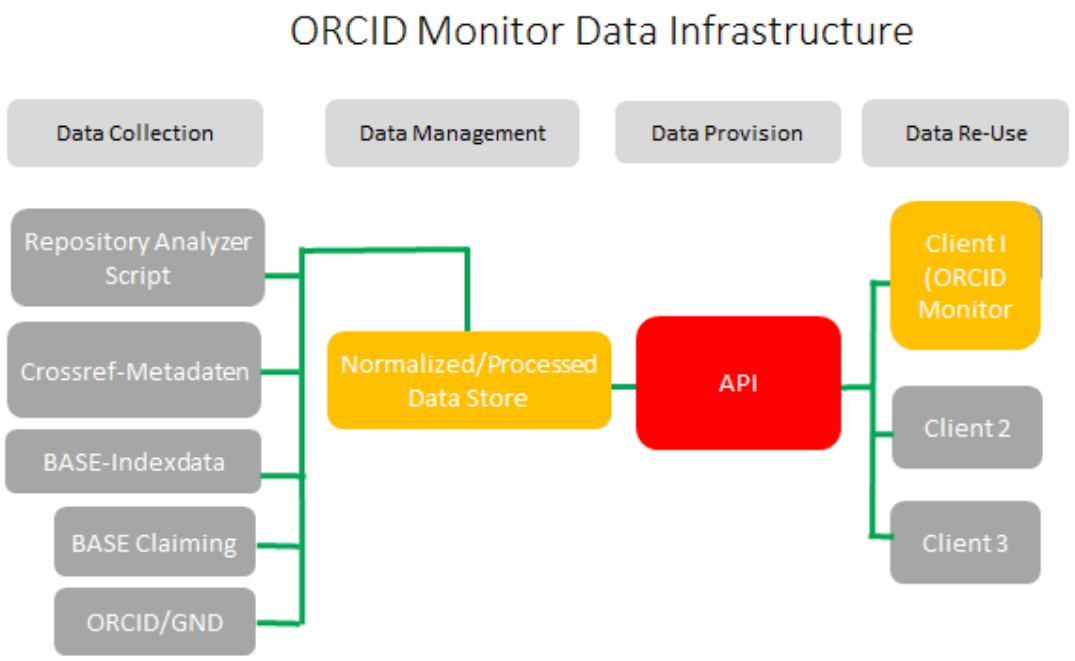

Abb. 1: ORCID DE Monitor Dateninfrastruktur

Mit diesem breit angelegten Ansatz wird eine uneingeschränkte, flexible und individuelle Nachnutzung der ermittelten Zahlen möglich. Durch die prototypische Implementierung einer Endnutzer*innenoberfläche wird auch interessierten Nachnutzer*innen ohne technische Implementierungsmöglichkeiten eine nutzbare Analyseplattform angeboten. Diese Benutzeroberfläche enthält dynamische Abfrage- und Einstellmöglichkeiten geeigneter Parameter, um flexible Auswertungen zu ermöglichen.

Die Monitorfunktionen sollen dabei Auswertungen und Daten liefern, die den Ist-Zustand der ORCID-Nutzung in Verbindung zu vorliegenden Faktoren beschreiben und zugleich auf Basis der über einen längeren Zeitraum vorhandenen Daten Entwicklungen analysieren lässt. Gleichzeitig soll durch die Bereitstellung der Vergleichszahlen (auch diese mit Darstellung der zeitlichen Entwicklungen) auf Länderebene Rückschlüsse für die Wirksamkeit anderer Strategien in anderen

Publikationscommunities unter Berücksichtigung der entsprechenden Umgebungsvariablen ermöglichen.

Damit kann ein Interessentenkreis mit Informationen versorgt werden, die Aufschlüsse über Ursachen und Konstellationen, aber auch Nutzen und Aufwand von Strategien zur weiteren Verbesserung der Verbreitung vorbereiten können. Für diese Zahlen von strategischer Bedeutung können Interessenten im Bereich Publikationsdienste wie Bibliotheken, universitäre und forschungspolitische Entscheidungsträger und auch Repository Manager in Frage kommen.

Die angestrebte Implementierung basiert auf der aus verschiedenen Kontexten vorhandenen Expertise der UB Bielefeld im Bereich Data Science und Visualisierung (s. insbes. [3]). Es werden insbesondere JQuery und bootstrap zur Realisierung der Dashboard-Oberfläche im HTML-Rahmen verwendet. Für die Datenvisualisierung mit üblichen Diagramm-Darstellungen wird d3js als Javascript Framework und für die Erstellung von Karten Google Chart verwendet, letzteres um speziell länder-basierte Vergleiche zu unterstützen (Näheres zu den Grundlagen s. [3]). Konkret muss eine Startseite mit Einstellmöglichkeiten entwickelt werden, die die Auswertungsparameter in einer komfortablen, assoziativ verständlichen Weise für Endnutzer*innen anbietet. Angedacht ist zudem als Add-on zum Projektantrag eine Exportschnittstelle, um die Daten nachnutzbar bereitzustellen.

Referenzen:

[1] Bertelmann, R., Cruse, P., Niggemann, E., Pieper, D., Sens, I., Burger, M., Dasler, R., Dreyer, B., Elger, K., Fenner, M., Hagemann-Wilholt, S., Hartmann, S., Höhnow, T., Kett, J., Pampel, H., Pietsch, C., Schirrwagen, J., Summann, F. (2019). ORCID DE 2 – Konsolidierung der ORCID-Informationsinfrastruktur in Deutschland.

[https://doi.org/10.2312/lis.20.01]

[2] Dreyer, B., Hagemann-Wilholt, S., Vierkant, P., Strecker, D., Glagla-Dietz, S., Summann, F., Pampel, H., Burger, M. (2019). Die Rolle der ORCID iD in der Wissenschaftskommunikation: Der Beitrag des ORCID-Deutschland-Konsortiums und das ORCID-DE-Projekt. *ABI Technik*, 39(2), 112-121.

[https://doi.org/10.1515/abitech-2019-2004]

[3] Summann, F., Czerniak, A., Schirrwagen, J., Pieper, D. (2020). Data Science Tools for Monitoring the Global Repository Eco-System and its Lines of Evolution. *Publications*, *8*(2), 35.

[https://www.mdpi.com/2304-6775/8/2/35]

[4] Summann, F. (2016). Die Verwendung von Autorenidentifikatoren in wissenschaftlichen Repositorien: Ansätze, konkrete Umsetzungen und Herausforderungen. Presented at the 105. Deutscher Bibliothekartag in Leipzig 2016 = 6. Bibliothekskongress, Leipzig.

[https://pub.uni-bielefeld.de/download/2907846/2907848/restored_btag_2016_orcid.pdf]

[5] Fenner, M., Hartmann, S., Müller, U., Pampel, H., Reimer, T., Scholze, F., Summann, F. (2016). Autorenidentifikation für wissenschaftliche Publikationen. Bericht über den Workshop der DINI-AG Elektronisches Publizieren auf dem 6. Bibliothekskongress. *o-bib*, 3(4), 286-293.

[https://doi.org/10.5282/o-bib/2016H4S286-293]

[6] Glagla-Dietz, S., Habermann, N. (2020). Standardnummern für Personen - Qualitätsverbesserung durch das Zusammenspiel intellektueller und maschineller Formalerschließung. *Dialog mit Bibliotheken*, 32(2), 20-25.

[https://nbn-resolving.org/urn:nbn:de:101-2020062250]

[7] Blogbeitrag [https://www.orcid-de.org/100000-gnd-orcid-verknuepft]

[8] Blogbeiträge [https://www.orcid-de.org/orcid-claiming-gnd] und [https://www.orcid-de.org/claiming-service-gnd-orcid]

Anhang 1 Beispiele:

Die Liste der folgenden Beispiele zeigt zunächst zwei Graphiken (Abb. 1 und 2), die zu verschiedenen Anlässen im Projektrahmen (Vorträge bei Veranstaltungen, ORCID Workshops) aufbereitet worden sind. Sie zeigen erste Auswertungsaspekte. Abb. 3 und 4 zeigt erste Prototypen der ORCID DE Monitorentwicklung, die in der Entwurfsphase eine Diskussionsgrundlage gebildet haben.

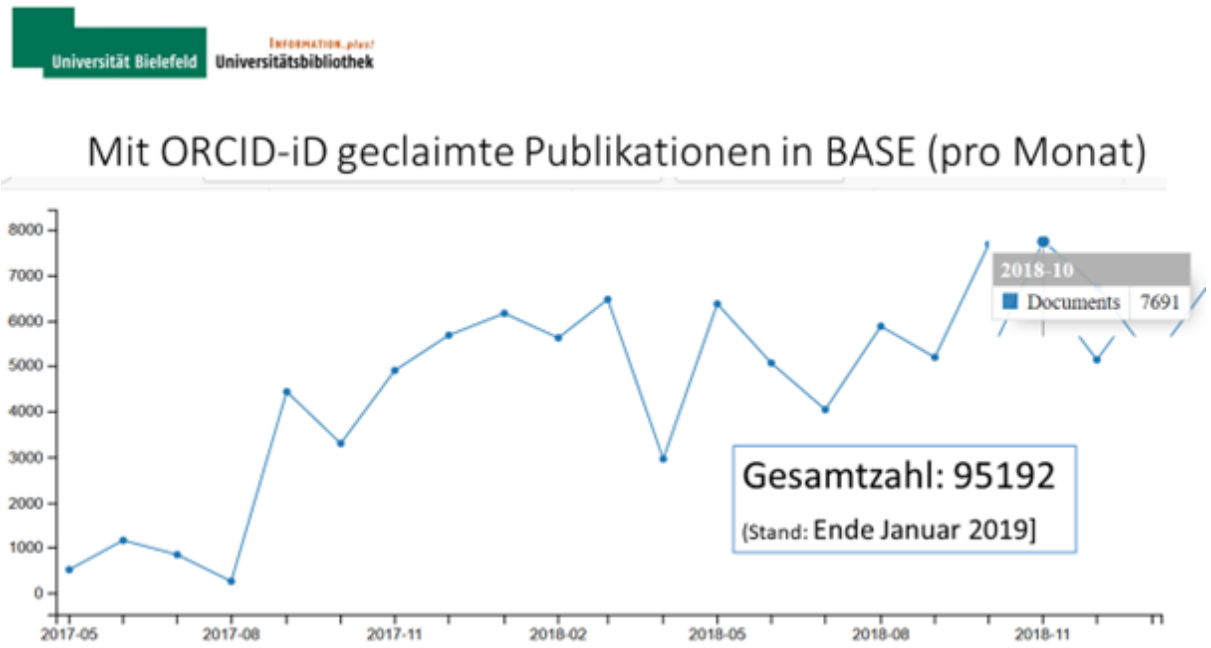

Abb. 1: Via ORCID iD geclaimte Publikationen bei BASE (2016 bis 2019)

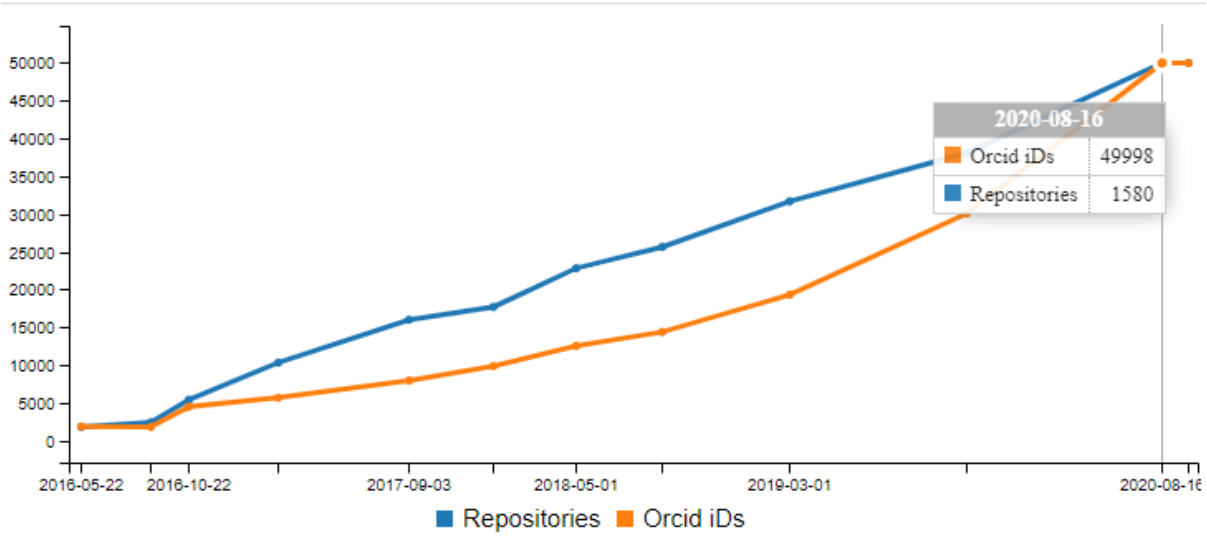

Abb. 2: Nachgewiesene ORCID iDs in Repository-Stichproben (2016 bis 2020)

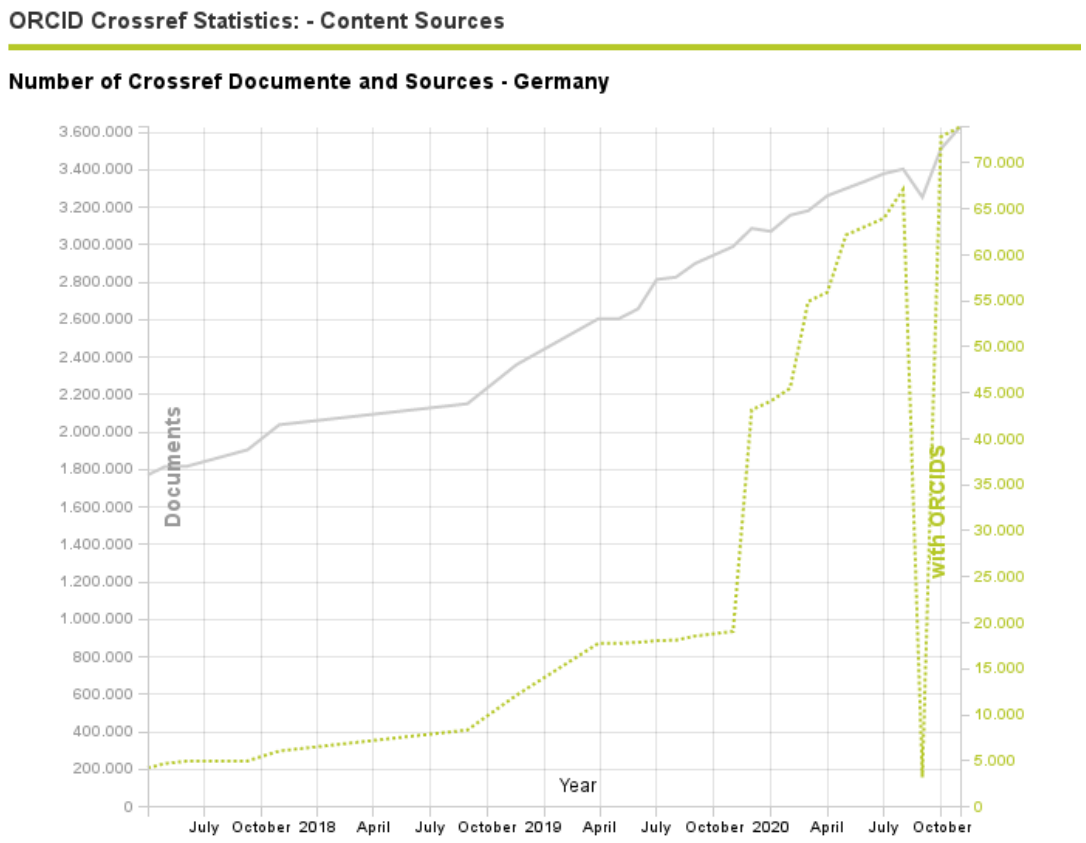

Abb. 3: Entwicklung des Vorkommens von ORCID iDs in den Crossref-Metadaten - Deutschland

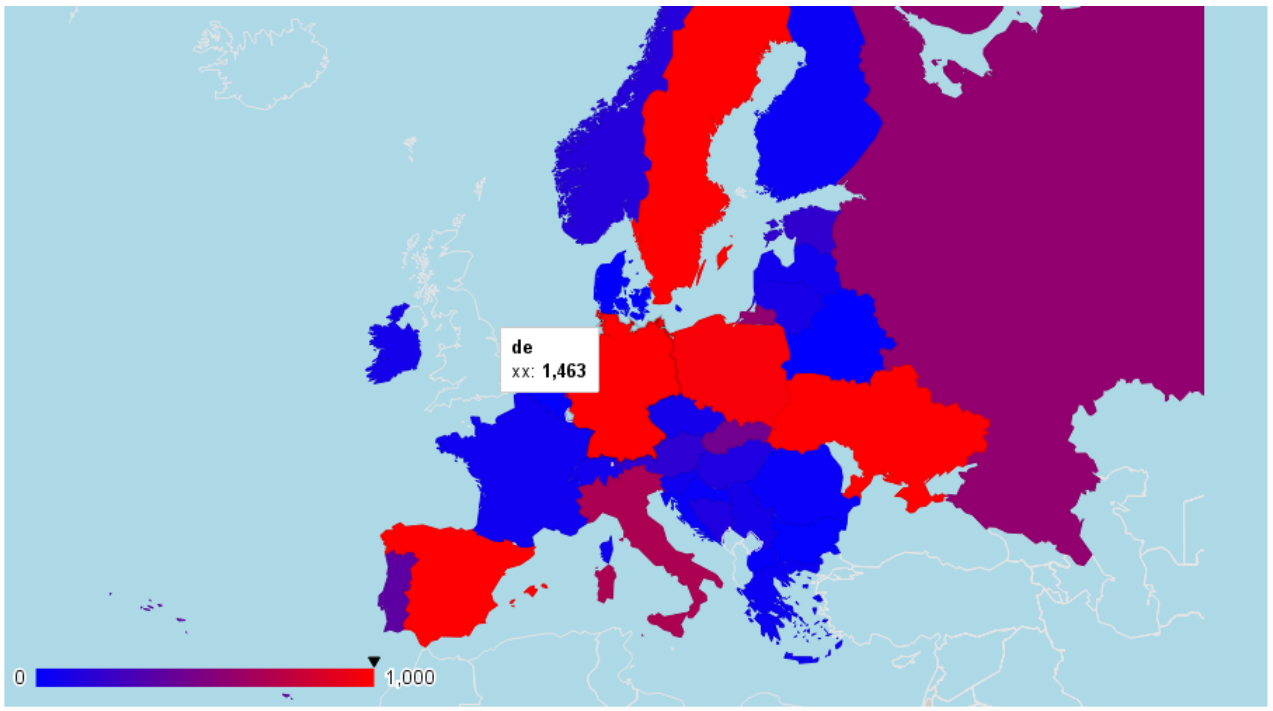

Abb. 4: Europa-Karte der Anzahl nachgewiesener Dokumente mit ORCID iDs