Discovery und Geodaten

Neue explorative Wege mit Graphen und Vektoren

1. Ziel: Auffindbarkeit und Kontextualisierung im Discovery

Discovery-Systeme wie das an der ETH-Bibliothek Zürich verwendete Primo VE [https://eth.swisscovery.slsp.ch/discovery/search?vid=41SLSP_ETH:ETH] bieten den Vorteil, dass Metadaten aus verschiedenen Quellen mit nur einer Suchanfrage durchsucht werden können. Die Ergebnisse der Suchanfrage stammen nicht nur aus dem eigentlichen Bibliothekskatalog (Alma), sondern auch aus Archivbeständen, Sammlungen, Bilddatenbanken, institutionellen Repositorien, Artikeldatenbanken und anderen Quellen.[1]

Die Suchergebnisse werden in einer flachen Ergebnisliste dargestellt: Die Liste ist linear und ohne Hierarchie. Unterschiedliche Typen oder Kategorien von Ergebnissen sind nicht ersichtlich. Dies erschwert die Filterung relevanter Informationen bei einer großen Anzahl von Ergebnissen. Wie kann dies für die Nutzenden vereinfacht werden? Gibt es neben den üblichen Facetten noch andere, effektivere Möglichkeiten?

Andersherum gefragt: An der ETH-Bibliothek gibt es viele Ressourcen. Wie lassen sich diese besser auffinden? Und wie können die Ressourcen untereinander besser verknüpft werden? Es gibt zahlreiche Möglichkeiten von Verknüpfungen: räumliche, zeitliche, inhaltliche und personenbezogene. Jede dieser Verknüpfungen kann ein Weg sein, die Ressource zu finden oder in neue Kontexte zu setzen. Wie können diese Verknüpfungen im Discovery besser sichtbar gemacht werden?

2. Projekt „Geodaten im Graph“

2.1 Motivation

Das Projekt „Geodaten im Graph“ (Januar 2023 bis Mai 2024) griff diese Fragen und bereits bestehende Lösungsansätze auf, um eine einheitliche Lösung zu finden.

Zu Beginn des Projektes untersuchte die an der ETH-Bibliothek Zürich arbeitende Gruppe „Rara und Karten“ ihre Bestände auf räumliche Angaben in den Inhalten und Metadaten. Die Daten liegen in unterschiedlicher Form vor:

1 Dieser Beitrag bezieht sich auf den gleichnamigen Vortrag des Autors am 06.06.2024 auf der 112. BiblioCon 2024 in Hamburg.

- e-maps, ein Angebot elektronischer Karten: es umfasst moderne, aber auch digitalisierte und georeferenzierte historischen Karten aus dem eigenen Bestand.
- ETHorama [https://ethorama.library.ethz.ch]: ausgewählte digitalisierte Inhalte werden auf einer virtuellen Karte mit bestimmten Orten verknüpft (auch „Geotagging“ genannt). Ausserdem sind historische und georeferenzierte Reiseberichte und Dossiers dargestellt.
- Digitalisierte Karten auf der Plattform für alte Drucke „e-rara“ [https://www.e-rara.ch/]
- Randkoordinaten von Karten in den Katalogaufnahmen

Diese Daten unterscheiden sich stark in Format und Struktur. Ist es möglich, die genannten unterschiedlichen Bestände geografisch durchsuchbar zu machen? Ein typischer Anwendungsfall wäre, bei einer Suchanfrage mit einer Punktkoordinate (Breitengrad, Längengrad) alle Ressourcen anzuzeigen, die einem Ort in der Nähe dieses Punktes zugeordnet sind. Alternativ könnte man herausfinden, in welchen Karten dieser Punkt liegt.

2.2 Graphdatenbank

Doch wie lassen sich unterschiedliche Geodaten so miteinander in Beziehung setzen, dass sie möglichst flexibel und zugleich explorativ auswertbar sind?

Die Wahl fiel auf eine Graphdatenbank. Eine Graphdatenbank besteht aus Knoten und Kanten, wobei Kanten die Knoten verbinden und den Typ dieser Verbindung definieren. Zum Beispiel führt von einem Knoten „EMap“ eine Kante des Types „DESCRIBES“ zu einem Knoten des Types „Place“. Oder von einem Knoten des Types „Contributor“ führen Kanten „HAS-CONTRIBUTED“ zu Knoten des Typs „ERaraItem“ oder „EMap“ (vgl. Abb. 1).

In einer Graphdatenbank bilden diese Tripel aus zwei Knoten und einer Kante die Grundstruktur der Daten. In diesem Knoten-Netzwerk kann man von jedem Knoten zu jedem anderen Knoten gelangen, es fragt sich nur, über wieviel Kanten man gehen muss. Hier zeigt sich der explorative Charakter dieser Datenstruktur.

Wenn Datensätze im Marc-Format in die Graphdatenbank eingelesen werden (z.B. Datensätze für Karten) so wird beim Einlesen in den Graphen dieser Marc-Datensatz zerlegt. Es resultiert nicht nur ein Knoten für die Karte. Das Marc-Feld für Kontributoren z.B. wird in einen oder mehreren Knoten (pro Kontributor) umgewandelt oder das Marc-Feld für Schlagwörter wird ebenso zu eigenen Knoten geformt. Diese werden mit einer „HAS-CONTRIBUTED“- bzw. „HAS-TOPIC“-Kante mit dem Karten-Knoten verbunden.

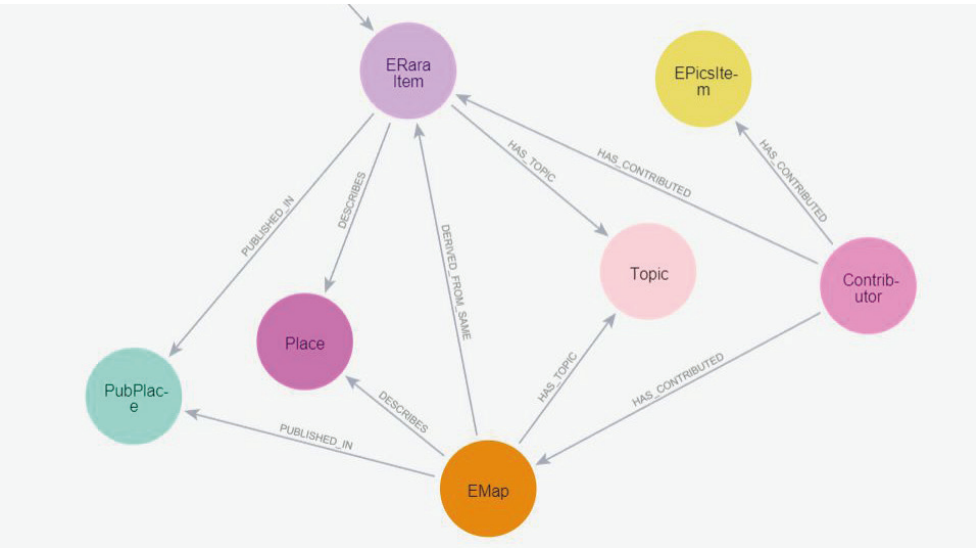

Abb. 1: Als Graphen strukturierte Geodaten

Gibt es GND-IDs zu Kontributoren oder Orten in der Katalogaufnahme, findet eine Anreicherung mit weiteren Identifikatoren und Informationen aus Wikidata (Orte) und VIAF (Personen) statt. Im Falle der Orte werden aus Wikidata neben Identifikatoren auch die Koordinaten des Ortes eingelesen. Indem einzelne Felder bzw. Unterfelder aus Marc-Datensätzen zu eigenen Knoten werden und die Beziehung im Typ der Kante ausgedrückt wird, werden aus flachen Datenstrukturen dreidimensionale Strukturen geformt. Künftig werden weitere ortsbezogene Daten eingelesen werden.

2.3 Abfrage über API

Wie kann diese flexible Datenstruktur genutzt werden?

Die ETH-Bibliothek bietet verschiedene API-Zugriffe in ihrem Developer Portal [https://developer.library.ethz.ch/] an. Hier gibt es zwei unterschiedliche Zugänge zu der Datenbank mit den Geodaten.

Die REST-API stellt verschiedene Endpunkte für die Recherche bereit, beispielsweise für Karten, Orte, Historische Reisen oder Topics. Diese Endpunkte lassen sich entweder mit einem Suchbegriff oder mit den Koordinaten eines Punktes abfragen. Für jeden Endpunkt ist genau festgelegt, welche Ergebnisse zurückkommen. Die Ergebnisse erhält man im GeoJSON-Format, das in anderen Tools (OpenStreetMap/Leaflet, geojson.io usw.) weiterverwendet werden kann.

Eine flexiblere Abfragemöglichkeit bietet GraphQL. Hier lässt sich in der Abfrage genauer bestimmen, welche Ergebnisse und welche Felder der Ergebnisse in der Antwort stehen. In der Abfrage können auch die gewünschten Kanten miteinbezogen werden. Dieser Typ der Abfrage ist explorativer, verlangt aber auch Vorkenntnisse im Umgang mit GraphQL.

3. Mehr Exploration im Discovery

3.1 Ortsseite

Wie lassen sich diese Abfragemöglichkeiten in unserem Discovery-System nutzen? In Discovery-Systemen sind Entitäten wie Personen oder Orte nur schwach repräsentiert. Deswegen gibt es eine neue Struktur im Discovery, die sogenannten „Ortsseiten“. Ortsseiten führen Informationen, Links und Ressourcen zu einem Ort innerhalb des Discovery-Systems zusammen.

Ortsseiten bieten u.a. ausgewählte Ressourcen zu dem jeweiligen Ort an. Es gibt zwei Arten von kuratierten Listen zum Ort:

- Ressourcen zu einem geografischen Schlagwort, das in der Katalogaufnahme vergeben wurde
- Ressourcen, die in ETHorama [https://ethorama.library.ethz.ch] einem Ort zugeordnet sind

Neben den kuratierten Listen zeigt die Ortsseite Links zu dem jeweiligen Ort in anderen Portalen (GND, GeoNames, Wikidata, archINFORM, Historisches Lexikon der Schweiz usw.) an, die in der Regel über den Abgleich gemeinsamer Identifikatoren maschinell generiert werden. Die ebenfalls verknüpften Themensammlungen und historische Reisen aus ETHorama eröffnen historische und systematische Kontexte (vgl. Abb. 2).

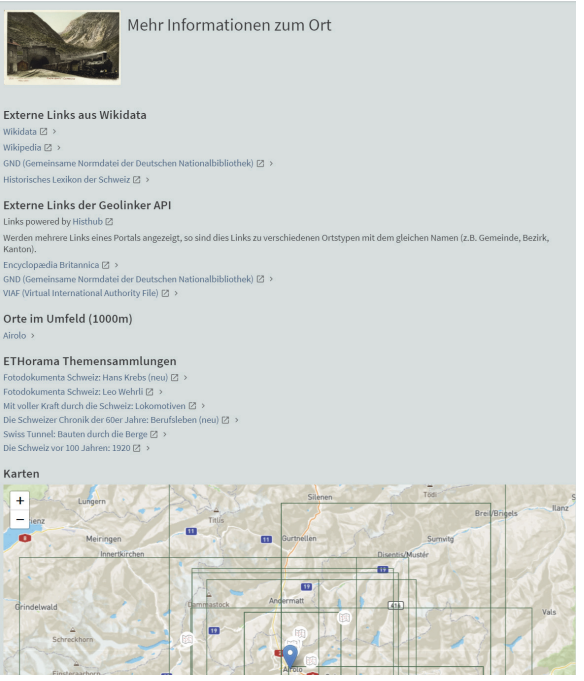

Abb. 2: Ortsseite mit mehr Informationen zum Ort

Außerdem sieht man auf einer Karte die Umrisse der verschiedenen online verfügbaren Karten, auf denen dieser Ort liegt. Durch einen Klick lässt sich die Online-Karte aufrufen.

Wie gelangt man zu einer Ortsseite?

Nutzende des Discovery-Systems gelangen auf die Ortsseite, wenn eine Karte, ein Druck oder eine andere Art von Ressource, sei es im entsprechenden Marc-Feld der Katalogaufnahme oder über ETHorama, eine Zuordnung zu dem Ort hat. Dann wird bei dieser Ressource ein Link zu der Ortsseite angezeigt. Auf diese Weise sind nach einem Klick auf den Link auch andere Ressourcen und Informationen zu dem Ort zu sehen (vgl. Abb. 3).

Abb. 3: Der geografische Bezug verlinkt ein Dokument mit einer Ortsseite

Die Verwendung der Graphdatenbank mit Geodaten im Discovery schafft also mehr explorative Möglichkeiten im Discovery.

3.2 Personenseite

Vergleichbares bietet unser Discovery-System für den Entitätstyp „Person“ an, der für Nutzende oft von großem Interesse ist, im Discovery aber nur eine schwache Repräsentation findet.

Mit Hilfe von Wikidata und verschiedenen Linked-Data-Anwendungen werden Personenseiten erzeugt. Die Personenseite ist wie die Ortsseite in zwei Abschnitte unterteilt:

Im ersten Abschnitt sind Ressourcen dieser Person gelistet, standardmäßig nach GND-ID selektiert. Der Nachteil einer solchen präzisen Selektion ist oft, dass nicht alle Ressourcen dieser Person gefunden werden. Deswegen werden mehrere Suchoptionen als Links angeboten, die zunehmend mehr, aber auch möglicherweise unpassende Resultate liefern (Precision/Recall-Problematik).

Im zweiten Abschnitt werden Informationen zu dieser Person dargestellt: biografische Informationen, Links zu dieser Person in anderen Portalen, Links in Archive, Forscherprofile und Links zu Personenseiten von Student*innen, Professor*innen, Verwandten usw. (vgl. Abb.4). Letzteres basiert auf Informationen aus der GND und Wikidata. Zum Beispiel ist auf der Personenseite von Hermann von Reichenau ein Link zu der Personenseite von Berthold von Reichenau zu sehen. Klickt man diesen Link, erschließt man sich die Ressourcen dieses Schülers von Hermann.

Durch die Teilnahme an Metagrid [https://metagrid.ch/] können auch spezifisch schweizerische Quellen wie z.B. das „Historische Lexikon der Schweiz“ oder die „Diplomatischen Dokumente der Schweiz“ integriert werden. Metagrid vernetzt eine Vielzahl sozial- und geisteswissenschaftlicher Ressourcen.

Wie gelangt man zu einer Personenseite? Immer wenn die Metadaten einer Ressource eine GND-ID zu einer Person enthalten, so wird ein Link zu der zugehörigen Personenseite erzeugt. Dank Metagrid werden auch in mehreren anderen Portalen Links zu unseren Personenseiten angeboten.

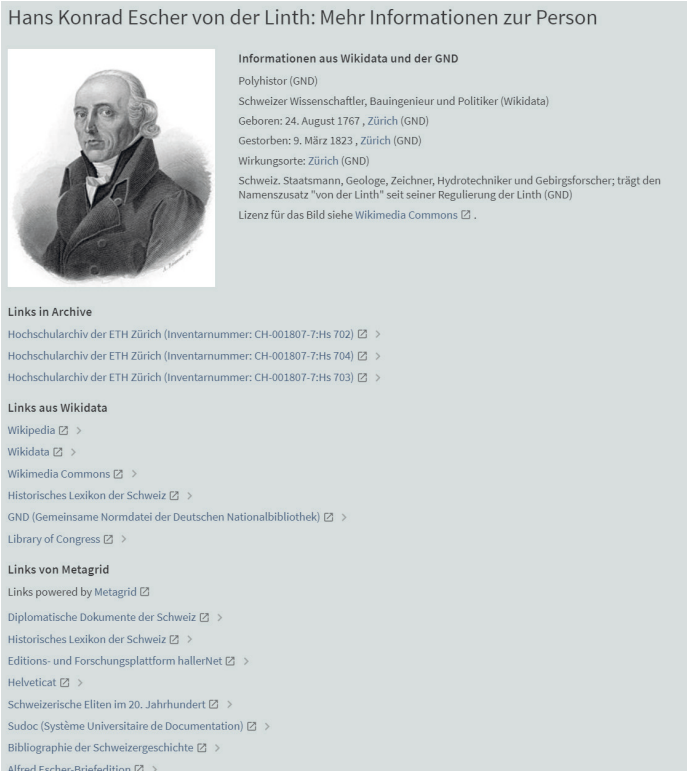

Abb. 4: Personenseite mit mehr Informationen zur Person

4. Angebote zur Sucheingabe

4.1 Orte und Personen

Man gelangt zu den Orts- oder Personenseiten, indem die entsprechenden Links auf den Detailseiten angeklickt werden. Auf diese Weise sind Ressourcen im Discovery über die Kategorien von „Ort“ und „Person“ mit anderen Ressourcen und Informationen verknüpft.

Aber wäre es nicht auch hilfreich, direkt nach der Eingabe eines Suchbegriffs Angebote zu Personen und Orten zu erhalten? Könnte es nicht parallel zur Ergebnisliste Angebote zu Personen und Orten geben? Lässt sich der Suchbegriff daraufhin überprüfen, ob ein Ort oder eine Person gemeint sein könnte? (vgl. Abb. 5)

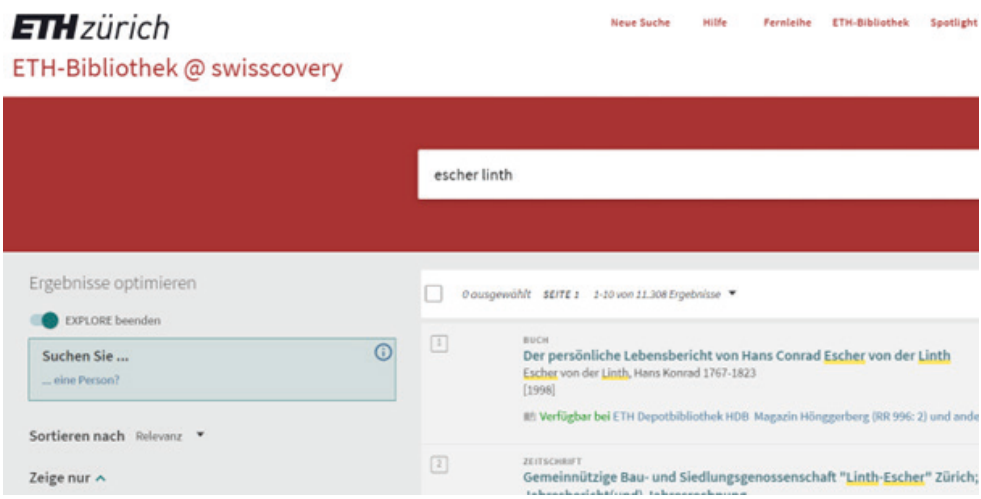

Abb. 5: Angebot parallel zur Ergebnisliste

Wird z.B. nach „escher linth“ gesucht, so besteht die Ergebnisliste aus Ressourcen zu Arnold Escher von der Linth, Hans Conrad Escher von der Linth und anderen Personen. Effektiver ist es so: parallel zur Ergebnisliste wird eine Liste von Personen in Form kleiner Personenkarten dargestellt, in diesem Fall von Arnold Escher und Hans Conrad Escher. Die Informationen und das Bild der Personenkarte helfen bei der Auswahl (vgl. Abb. 6). Klickt man auf die Karte von Hans Conrad Escher, kommt man zu dessen Personenseite, auf der standardmäßig die nach GND-ID selektierten Ressourcen zu sehen sind.

Abb. 6: Zum Suchbegriff passende Personen

Einen weiteren Vorteil bietet die Personenliste in der Suche nach Namensvarianten. Die Suche nach dem chinesischen Mathematiker „Wen-tsin Wu“ führt nur zu sehr wenigen Ergebnissen. Auf einer Personenkarte wird „Wu Wenjun“ vorgeschlagen. Durch Auswahl dieser Option gelangt man zu wesentlich mehr Ressourcen.

Ähnlich verhält es sich bei der Suche nach Orten: Suche man z.B. nach „Gotthard“ werden verschiedene Orte auf einer Karte angeboten, z.B. „Gotthardmassiv“, „Gotthardpass“, „Gotthardtunnel“ usw. Durch eine Auswahl gelangt man zu der jeweiligen Ortsseite.

Was steht dahinter? Zur Erzeugung der Personenliste wird mit dem Suchbegriff der Wikidata Query Service befragt. In Wikidata sind die Personen unter anderem mit sehr vielen, auch internationalen Namensvarianten enthalten.

Die Kartenansicht der angebotenen Orte basiert auf einer Abfrage der Graphdatenbank, also auf Orten aus (bislang) ETHorama, e-maps, e-rara oder dem Katalog (Alma).

Das Angebot von Personen und Orten bedeutet, dass der Schritt vom „String“ (Suchbegriff) zum „Thing“ (Objekt, Person oder Ort) vollzogen werden kann. Durch die Auswahl einer Person oder eines Ortes ist entschieden, um welches Objekt es geht. Zu diesem Objekt können dann weitere Informationen und Links hinzugefügt werden, da es eindeutig ist.

4.2 Themen

Auch die thematische Suche bringt oft nicht die gesuchten Ergebnisse.

Seit langem nutzen Bibliotheken Schlagwort-Kataloge für die thematische Suche. Doch wie gelangt man an die passenden Schlagworte? Die auf der Ergebnisliste basierenden Schlagwörter in den Facetten sind oft unzureichend. Eine gängige Praxis vieler Nutzender ist es deshalb, Detailseiten von Ressourcen durchzugehen und dort die passenden Schlagwörter einzusammeln.

Es gibt einen direkteren Ansatz: Nach der Sucheingabe werden semantisch ähnliche Schlagwörter vorgeschlagen (vgl. Abb. 7). In den Metadaten der vorgeschlagenen Schlagwörter muss der Suchbegriff nicht vorkommen. Eine semantische Ähnlichkeit ist genug. Wird beispielsweise nach „Wasserkraft“ gesucht, werden unter anderem „Theorie des Wasserbaus“ oder „Wasserräder (hydraulische Energie)“ vorgeschlagen.

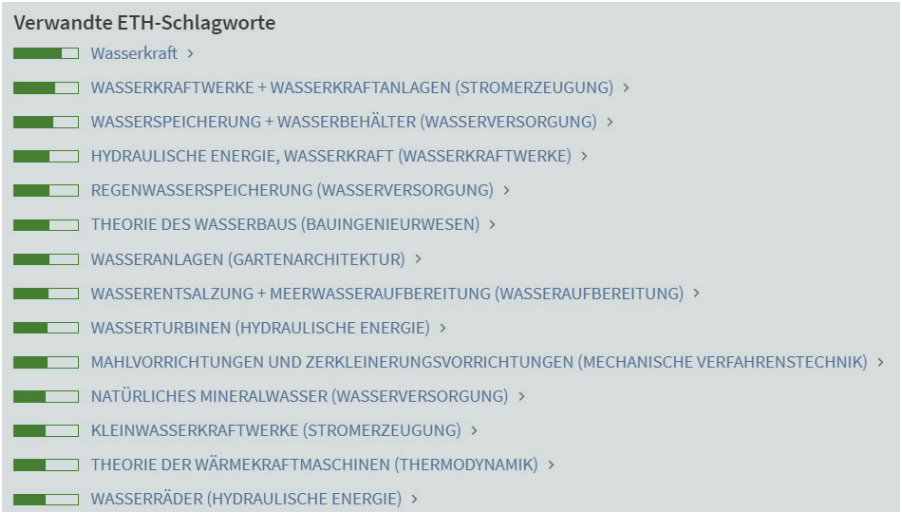

Abb. 7: Verwandte ETH-Schlagworte nach der Sucheingabe

Diese Ähnlichkeits-Suche basiert auf einer KI-gestützten Umwandlung der Metadaten des ETH-Schlagwort-Kataloges in Vektoren. Jeder Vektor ist eine Aneinanderreihung vieler Zahlen, die die semantische Bedeutung des Schlagwortes repräsentieren. Alle Vektoren sind in einer spezialisierten Vektordatenbank gespeichert. Ähnliche Schlagworte haben Vektoren, die nahe beieinander liegen.

Auch der Suchbegriff wird in einen Vektor verwandelt und die Vektordatenbank gibt hierfür ähnliche Vektoren (Schlagworte) zurück. Eine Auswahl eines der ähnlichen Schlagworte führt zur erweiterten Suche, die im Schlagwortfeld nach dem Schlagwort sucht. Falls sich dieser Ansatz in der Praxis als hilfreich herausstellen sollte, wird die Vektordatenbank über den Schlagwort-Katalog der ETH hinaus erweitert werden.

Auch im Fall der Schlagworte kommt die langjährige sorgfältige Arbeit der Bibliothekar*innen in neuer Weise zur Geltung. Metadaten und Schlagwörter erfahren durch den Einsatz von Techniken wie Graphen und Vektoren also eine Aufwertung. Diese neueren Techniken brauchen eine solide Datenbasis, um Nutzenden eine bessere Erfahrung im Discovery zu ermöglichen.

Bernd Uttenweiler, ETH-Bibliothek Zürich, https://orcid.org/0009-0006-8613-8183

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/6077

Dieses Werk steht unter der Lizenz Creative Commons Namensnennung 4.0 International.