

Die Integration von Daten in das Verbundfindmittel: Informationen zu Thema Harvesting

Im Rahmen des Projektes „Ausbau des Netzwerks SED-Archivgut zu einer Referenzanwendung für ein Archivportal Deutschland“ können die am SED- und FDGB-Netzwerk teilnehmenden Archive ausgewählte Findmittel aus ihren Beständen unter der einheitlichen Oberfläche des Verbundfindmittels veröffentlichen und für eine archivübergreifende Recherche bereitstellen. Hierzu müssen die Beständeübersichten und Findbücher auf den zentralen Server übertragen werden.

Im Projektantrag wurden zwei Möglichkeiten für die Integration von Daten in das zentrale Verbundfindmittel angesprochen. Die eine Möglichkeit ist das direkte Hochladen der Dateien auf den zentralen Server (mit WebDAV oder ftp-Upload). Die andere Möglichkeit wäre das regelmäßige Abholen der Daten auf den dezentralen Servern der Archive (Harvesting). Beide Wege werden im Folgenden dargestellt und abgewogen.

1. Hochladen der Daten auf den zentralen Server

Das Hochladen geschieht aktiv durch das Archiv und kann jederzeit vorgenommen werden. Das Ergebnis ist kurzfristig, normalerweise am folgenden Tag, sichtbar. Durch das Projekt wird ein Werkzeug bereitgestellt, mit dem die Daten für das Hochladen ausgewählt und überspielt werden. Wenn sie auf dem zentralen Server angekommen sind, wird ein Indexierungsvorgang angestoßen, der die Daten für die Suche vorbereitet. In gleicher Weise können die Dateien wieder gelöscht werden. Sie werden dann von dem zentralen Server entfernt und können nicht mehr durchsucht und angezeigt werden.

Dieser Weg bietet mit dem manuellen Einspielen bzw. Löschen der Findmittel eine direkt nachvollziehbare Einflussnahme durch das bereitstellende Archiv über eine Internetverbindung, ist jedoch auch mit einer aktiven Bedienung und Beaufsichtigung des Werkzeugs verbunden. Dieses Verfahren ist bereits jetzt implementiert.

2. Abholung der Daten auf einem dezentralen Server durch einen Harvester

Alternativ könnten die Daten mit Hilfe der sogenannten Harvesting-Technologie übertragen werden. Hierbei läuft nach der Einrichtung und Konfiguration des Systems die weitere Synchronisierung mit dem Verbundfindmittel automatisch ab. Die Bearbeitung und Erstellung der EAD-Findmittel selbst erfolgt weiterhin über das Konvertierungswerkzeug MDEX. Die mit diesem für das Verbundfindmittel aufbereiteten Daten werden auf dem Webserver des Partnerarchivs abgelegt, auf dem zusätzlich eine Software für die Harvester-Komponente installiert werden muss. Diese ist kostenlos bzw. unter Opensource-Lizenz frei verfügbar und wird vom Bundesarchiv bereitgestellt. Die Harvester-Komponente registriert sich automatisch beim Verbundfindmittel und stellt ein Verzeichnis aller zur Verfügung stehenden Findmittel des Partnerarchivs bereit.

Der Harvester des Verbundfindmittels besucht in regelmäßigen Abständen alle registrierten Webserver der Partnerarchive und analysiert die dort bereitgestellten Findmittel. Sie werden dann mit den im Verbundfindmittel gespeicherten Beständen abgeglichen. Dadurch werden vom Partnerarchiv neu bearbeitete und bereitgestellte Findmittel in das Verbundfindmittel hochgeladen oder vorhandene Findmittel aktualisiert und ein neuer Index erstellt. Sollte ein bereits veröffentlichtes Findmittel durch das Partnerarchiv auf dem eigenen Webserver gelöscht worden sein, wird es auch im Verbundfindmittel gelöscht und aus dem Suchindex entfernt. Da ein solcher Harvester die Webserveradressen, die er gespeichert hat, in festen Abständen besucht, muss dieser Termin abgewartet werden, bevor eine Veränderung im Verbundfindmittel feststellbar ist.

Als technische Voraussetzung für das Harvesting muss das Partnerarchiv einen Webserver bereitstellen, der permanent im Internet erreichbar ist. Dies setzt voraus, dass mit den IT-Verantwortlichen der jeweiligen Einrichtung der Zugriff z.B. hinsichtlich eventuell vorhande-

ner Sicherheitsvorrichtungen wie einer Firewall abgestimmt ist. Für den Webserver sind Microsoft Windows oder Linux als Betriebssystem erforderlich. Darüber hinaus müssen die durch das Bundesarchiv bereitgestellten Softwarekomponenten auf dem Webserver installiert werden.

Nach den bisherigen Rückmeldungen aus den beteiligten Archiven geht das Projektteam im Moment davon aus, dass die Realisierung des Harvestings in diesem Projekt nicht erforderlich ist und das eigene, aktive Hochladen besonders zu Beginn der Zusammenarbeit im Verbundfindmittel vorgezogen wird. Es erfordert weniger Voraussetzungen bei den Archiven selbst und die Wirkungen der Integration von Daten in das Verbundfindmittel sind schneller und direkter nachvollziehbar. Damit wird nicht ausgeschlossen, dass zu einem späteren Zeitpunkt im Routinebetrieb das Harvesting zusätzlich zum aktiven Hochladen ebenfalls angeboten werden wird.