Themis – Sicheres Vererben in der Cloud

Andrew Lindley[1] · Sebastian Pimminger[2] · Wolfgang Eibner[3] · Mihai Bartha[1]

[1]AIT Austrian Institute of Technology GmbH
{Andrew.Lindley, Mihai.Bartha}@ait.ac.at

[2]FH OÖ Forschungs & Entwicklungs GmbH
Sebastian.Pimminger@fh-hagenberg.at

[3]X-Net Services GmbH
we@x-net.at

Zusammenfassung

Vererben bedeutet teilen mit der Zukunft. Das Themis Projekt stellt eine sichere und skalierbare Plattformlösung bereit, die es dem Endanwender erlaubt, regelmäßig und automatisch persönliche Backups aus online Diensten wie Facebook, Moodle, E-Mail, Dropbox, dem Smartphone und Desktopquellen zu erzeugen. Themis gibt dem Benutzer die Kontrolle über seine Daten zurück und stellt darüber hinaus nützliche Anwendungen im Hinblick auf das private digitale Erbe bereit. Hierzu zählen etwa eine leistungsstarke Volltext- und Metadatensuche über den gesamten archivierten Datenbestand des Benutzers, innovative Formen der Datenaufbereitung über Raum und Zeit sowie die Möglichkeit, Datensätze individuell zu arrangieren und diese mit anderen Benutzern auf der Plattform sicher zu teilen, ohne hierbei die Kontrolle darüber zu verlieren. Diese Arbeit zeigt die Notwendigkeit eines privaten digitalen Archivs auf und präsentiert die Motivation und Umsetzung eines solchen Generationen-Backups in Form eines juristisch geregelten Vererbungs-Workflows. Dieser ermöglicht es gespeicherte Daten im Falle des Ablebens sicher an die Nachfahren weiterzuvererben. Privacy und Security sind notwendige Grundlagen eines solchen Dienstes. Anhand der beschriebenen Anwendungsfälle „Backup erstellen“, „Backup ausführen“ und „Backup teilen“ präsentieren wir die zentralen Punkte der Themis Sicherheitsarchitektur und innovative Forschungsergebnisse, wie etwa die Umsetzung des Index-Per-User Modells über Elasticsearch und die asynchrone Entkopplung der Datenindizierung.

1 Motivation und Ziele

Wir leben in einem Zeitalter der Informationsgesellschaft welches durch Daten geprägt ist. Ein Leben ohne Internet ist für die meisten unvorstellbar geworden - Information ist die Währung des 21 Jahrhunderts. Aktuelle Studien von IDC [GaRe12] zeigen, dass sich das Volumen an weltweit gespeicherten Daten alle zwei Jahre verdoppelt und ein Volumen von 40.000 Exabytes im Jahre 2020 erreichen wird. Es sind zwei Trends absehbar. Erstens, während sich früher Firmen und öffentliche Einrichtungen für die Datenerzeugung

hauptverantwortlich zeichneten, so zeigt die Studie eine klare Verschiebung zu Gunsten des Endanwenders. Bereits 2012 waren Konsumenten für 68% aller weltweit kreierten und verarbeiteten Daten verantwortlich. Zweitens ein Paradigmenwechsel hin zu Cloud Services. Geht man etwa ein Jahrzehnt zurück, so war der Speicherort unserer digitalen Fotos, Dokumente und anderer Artefakte der private Computer. Dies hat sich geändert. Der Trend zu Cloud Services ist ungebrochen. IDC geht davon aus, dass 2020 etwa 40% aller Daten in der Cloud gespeichert oder von Cloud Dienstleistern verarbeitet werden. Folgende relevante Themen- und Problemfelder lassen sich aus diese Entwicklungen ableiten:

- **Datenstreuung und Nachvollziehbarkeit.** Die Daten eines einzelnen Benutzers liegen verteilt über eine Vielzahl an verschiedenen Services, Provider und Dienstleister. Es ist nicht immer einfach für den Endanwender nachzuvollziehen wo seine Daten physisch gespeichert bzw. welche Services überhaupt in Anspruch genommen werden. Dieser Umstand ist kritisch, zieht man in Betracht, dass ein Benutzer typischerweise Daten über den Zeitraum von mehreren Jahren oder Jahrzehnten produziert. [DJRW10] zeigt, dass einer der Hauptgründe für Datenverlust neben verlorengegangenen Zugriffsberechtigungen und -mechanismen, schlichtweg das Vergessen des Benutzers darauf ist.
- **Serviceabschaltung und Accountstillegung.** Das Angebot an verfügbaren Cloud-Services unterliegt einem stetigen Wandel hinsichtlich der Marktgegebenheiten und -anforderungen. Plattformen, die heute noch im Fokus stehen, können morgen bereits obsolet sein. Unternehmensinsolvenzen, Portfoliobereinigungen und daraus resultierende Serviceabschaltungen sind keine Seltenheit, wie sich an den Beispielen Twitpic, Geocities, Friendster, Google Reader, etc. exemplarisch zeigt. Mangelnde Exportmechanismen, fehlende finanzielle Mittel oder rechtliche Rahmenbedingungen [Chau14] sind Gefahren, die Daten unerreichbar machen und ein Loch in unserem persönlichen digitalen Erbe hinterlassen können.
- **Datenexport und Archivierung.** Dem durchschnittlichen Benutzer fehlt die notwendige Fachkunde und Fertigkeit seine digitalen Daten zu archivieren. Studien [MaSh14] haben gezeigt, dass Benutzer nicht in der Lage sind, ein persönliches Archiv ihres digitalen Erbes zu pflegen, unabhängig davon, ob die Daten lokal oder in der Cloud gespeichert sind. Hierbei sind zugrundeliegende Fragestellungen der digitalen Langzeitarchivierung im Bereich der Bitstream-Preservation, wie Dateiformatkonvertierung und deren Qualitätssicherung, bereits außen vor gelassen. In vielen Fällen ist der Datenexport seitens des Serviceproviders nur eingeschränkt oder gar nicht möglich, da dies eine Möglichkeit darstellt, Kunden an die eigene Plattform zu binden.
- **Dateninterpretation im Kontext des Services.** Persönliche digitale Daten sind immer stärker im Kontext des Serviceanbieters zu betrachten. So hat beispielsweise ein Bild auf Facebook eine andere Bedeutung für den Anwender wie genau dasselbe Bild, das lokal auf der Festplatte gespeichert ist oder auf Instagram hochgeladen wurde [LMB+13]. Selbst wenn die Möglichkeit des Datenexports vorhanden ist, geht bei der Archivierung im Regelfall die Bandbreite an möglichen Aktionen (z.B. teilen, bearbeiten, kommentieren), die über ein solches Service mit den Daten verbunden ist, verloren.
- **Zusammengesetzte Entitäten.** Zusammengesetzte digitale Entitäten stellen eine für den Endanwender logische Sicht auf seinen digitalen Datenbestand hinsichtlich einer Anwendung dar. Es handelt sich hierbei um, je nach Fragestellung, unterschiedliche

Projektionen auf den zugrundeliegenden Inhalt. So zeigt Harper et. al [HLT+13], dass es nicht der Erwartungshaltung des Benutzers entspricht, Bilder aus Social-Media Quellen aus den zugehörigen sozialen Metadaten wie Kommentaren und Likes loszulösen. Das ORE Datenmodell [McNS09] stellt hierfür das Konzept der Resource Map (ReM) vor, welches es erlaubt, beliebig komplexe, zusammengesetzte digitale Entitäten zu modellieren.

- **Verschwommene Grenzen des Besitztums.** Benutzer verspüren einen weniger ausgeprägten Sinn von Besitztum und Kontrolle über Inhalte in der Cloud. Daten aus der Cloud zu beziehen und diese lokal zu speichern verstärkt das Gefühl von Eigentum und Besitz [OSHT12]. Die Grenzen des Eigentums sind jedoch diffus und nicht immer einfach für den Endanwender auszumachen, wie sich etwa bei Dateifreigaben in Dropbox zeigt. Anspruch auf Besitz kann auch über die eigenen Daten hinausgehen [MaSh11]. So kann etwa jemanden in einem Foto zu markieren dahingehend interpretiert werden, dass die Rechte an dem Bild teilweise auf den Benutzer ausgeweitet werden. Entscheidungen der eigentlichen Datenhalter den Inhalt zu löschen oder nicht länger zu teilen sind in der Flut an digitaler Information nur schwer zu bemerken und kaum zu revidieren. Datenverlust im Netz kann somit verschiedensten Ursprungs sein und ist nicht immer technischer Natur [DJRW10].

Durch den zunehmenden Anstieg an Daten und dem darin enthaltenen Wissen benötigen Benutzer neue Möglichkeiten und Wege diese fassen zu können. Benutzer sollen die Herrschaft über ihre digitalen Daten zurückgewinnen und nach eigenen Bedürfnissen verwalten. So gibt es einige Werkzeuge und Systeme, die diese Probleme zu lösen versuchen.

Im Bereich der Sicherung aus Online-Plattformen gibt es unzählige Werkzeuge und Möglichkeiten wie z.B. ArchiveFacebook, Gmail Backup Tool oder Backupify. Tools wie diese sind im Regelfall auf ein bis wenige Services ausgerichtet, in ihrer Erweiterbarkeit stark eingeschränkt und erstellen physische Backups ohne selbst zusätzliche Nutzen wie Suche, Indizierung oder dergleichen anzubieten. Zur sicheren Verwahrung von Backups existieren spezialisierte Dienste mit verschlüsselten Speicherlösungen. Populäre Vertreter dieser Kategorie sind z.B. Wuala und SpiderOak. Es gibt aber keine Möglichkeiten, ohne zusätzliches Tooling, Backups auf diese Dienste automatisiert durchzuführen. Auch muss die zugesicherte Sicherheit angezweifelt werden, denn Dienste wie z.B. Wuala veröffentlichen ihren Quellcode nicht, selbst nicht für verschlüsselungsrelevante Teile des Systems. Dienste wie die iOS und Android App Timeshop erlauben es, die Vergangenheit chronologisch in Form von Social Media Beiträgen und Smartphone-Fotos Revue passieren zu lassen, bieten jedoch keine Möglichkeiten, diese Daten zu archivieren.

Im betrieblichen und öffentlich-rechtlichen Umfeld existieren klare Regelungen und gesetzliche Vorgaben zu Archivierung. Aspekte der digitalen Langzeitarchivierung waren Bestandteil zahlreicher Forschungsprojekte [StPR11].

Lösungen aus diesem Umfeld, wie etwa zur skalierbaren und qualitätsgesicherten Datenmigration [SCM+13] haben aufgrund der technischen Komplexität und benötigten Expertise, bis auf wenige Ausnahmen [GSPR10], noch kaum Einzug in den Privatbereich gehalten. Was den Bereich geordnete und sichere Weitergabe sowie Vererben von Daten anbelangt, betritt man Neuland [GOFF13]. Dienste wie PasswordBox, bieten eine Erweiterung an, die es Erben erlaubt, im Ablebensfall auf die Zugangsdaten zu den darin verwalteten Accounts ihres Verstorbenen zuzugreifen. VitalLock stellt verschlüsselten Online-Speicher bereit, der es ermöglicht, wichtige Informationen für den Fall von Notfällen automatisiert zu versenden.

Die oben genannten Werkzeuge und Systeme bieten zwar für ihre Bereiche gute und adäquate Lösungen, haben aber einen entscheidenden Nachteil. Um Benutzer auch wirklich die Herrschaft über ihre Daten geben zu können, bedarf es einer integrierten Lösung der oben genannten Bereiche, angefangen von einer sicheren Speicherung der Daten über eine Datensicherung aus unterschiedlichen Plattformen und Diensten bis hin zu einer Indizierung und gesicherten Weitergabe der Daten. Außerdem muss diese Lösung sämtliche Stakeholder des Prozesses mit einbinden.

2 Die Themis Plattform

Themis[1] ist ein innovativer Dienst zur Sicherung, Verwaltung und Vererbung von persönlichen digitalen Daten aus Online-Services, dem Smartphone und dem PC, der auf die Bedürfnisse von Endanwendern maßgeschneidert ist. Dem Benutzer wird ein Service zur Verfügung gestellt, das ihm ermöglicht, die Herrschaft über digitale Daten zurückzugewinnen und Datenverlust vorzubeugen, gleichzeitig nützliche Funktionen aufweist und als Security-First Lösung umgesetzt wurde. Das technische Grundgerüst basiert hierbei auf den Entwicklungen des BackMeUp Projektes, in welchem zwischen 2011 und 2012 geforscht wurde und welches den Nachweis der Machbarkeit im Bereich persönlicher Webarchivierung angetreten ist. Ziel war es, Plugins für unterschiedliche Online-Dienste bereit zu stellen, einen Backup-Workflow zu entwickeln der Daten von diesen herunterlädt, in ein für den Benutzer nützliches und archivierbares Format überführt und in einer frei wählbaren Datensenke persistiert. Themis stellt eine sichere Plattform für den Benutzer zur Interaktion mit seinem persönlichen digitalen Erbe bereit, die BackMeUp um folgende Zielsetzungen erweitert:

- **Durchsuchbarkeit und Datenvisualisierung**. Der zunehmende Anstieg digitaler Daten erfordert neue Wege in der Suche und Visualisierung des darin enthaltenen Wissens. Die Bereitstellung eines Raum-, Zeit- und Rechte-affinen Index über alle Datenquellen hinweg ermöglicht das schnelle und zielgerichtete Auffinden gesicherter persönlicher digitaler Daten. Es ermöglicht beispielsweise Abfragen wie „zeige mir alle Kalendereinträge, Facebook Veranstaltungen, SMS und Word-Dateien vom 18. März 2015“ oder „finde Bilder die in der Nähe von Wien aufgenommen wurden“. Der Vorgang des Auffindens, die Neugruppierung von Datensätzen und deren Verlinkung über verschiedene Datenquellen hinweg ist ein essentieller Bestandteil des Projektes.
- **Vererben und Teilen des digitalen Nachlasses.** Archivierte Daten aus dem persönlichen digitalen Erbe können mit anderen Benutzern der Plattform geteilt und, im Falle des Ablebens, über einen juristisch geregelten Vererbungs-Workflow an die nächste Generation weitergegeben werden, ohne dabei die Kontrolle über die Datenherrschaft zu verlieren. Bereits zu Lebzeiten sollen dem Benutzer Möglichkeiten gegeben werden, um über seinen Nachlass an unterschiedliche Personengruppen zu bestimmen und diesen zu selektieren. Themis zeigt die Umsetzung eines solchen Prozesses mit der Österreichischen Notariatskammer und beschäftigt sich darüber hinaus mit den rechtlichen Rahmenbedingungen und wirtschaftlichen Faktoren.
- **Smartphones und weitere Datenquellen.** Mobile Endgeräte wie Smartphones und Tablets sind hochpersonalisierte Geräte, die eine Vielzahl an persönlichen Daten beinhalten, jedoch einem raschen Wandel unterworfen und in regelmäßigen Zyklen durch

[1] benannt nach der Göttin der Ordnung und Gerechtigkeit in der griechischen Mythologie.

neue Produkte abgelöst werden. Themis stellt Backupmöglichkeiten für den mobilen Bereich bereit, d.h. die Sicherung von Daten auf mobilen Geräten und der komfortable Abruf von gesicherten digitalen Inhalten über die Plattform. Die Sicherungsmöglichkeiten werden auf Home-NAS Systeme ausgedehnt, um die gesicherten digitalen Daten auch physisch vollständig zurück in die Einflusssphäre des Benutzers bringen zu können.

- **Security by Design.** Die Grundvoraussetzung zur Akzeptanz eines solchen Backup-Dienstes ist es, Datenmissbrauch vorzubeugen und die Sicherheit des persönlichen digitalen Erbes sicherzustellen. Themis verfolgt eine durchgängige Security by Design-Architektur über alle Komponenten des Systems, die es selbst dem Plattformbetreiber unmöglich macht, auf darin gespeicherte Daten seiner Benutzer zuzugreifen. [Siehe Kapitel: Architektur]
- **Skalierbarkeit.** Während BackMeUp als Proof-of-Concept Lösung ausgerichtet war, stellt Themis eine skalierbare Lösung an Backend-Komponenten bereit.
- **Bewusstseinsbildung.** Obwohl 2010 bereits eine Mehrheit der weltweit erzeugten digitalen Daten ihren Ursprung durch Einzelpersonen und nicht durch Organisationen fand, ist die Bewusstseinsbildung für die Wichtigkeit persönlicher digitaler Archive nur eingeschränkt vorhanden. [DJRW10] Vergessen ist ein wesentlicher Faktor der hierzu beiträgt. Über regelmäßige, automatisierte Backups, sowie der Möglichkeit des Backup Sharing, soll bereits zu Lebzeiten des Benutzers ein Bewusstsein für die Notwendigkeit eines persönlichen digitalen Archivs und die Akzeptanz der Plattform geschaffen werden.

3 Architektur

Themis ist als verteiltes System mit den Schwerpunkten Sicherheit und Skalierbarkeit konzipiert worden. Es gliedert sich in sechs größere Submodule, die in sich abgeschlossene Services darstellen und über wohldefinierte Schnittstellen miteinander kommunizieren. Abb. 1 gibt einen Überblick über den architektonischen Ansatz.

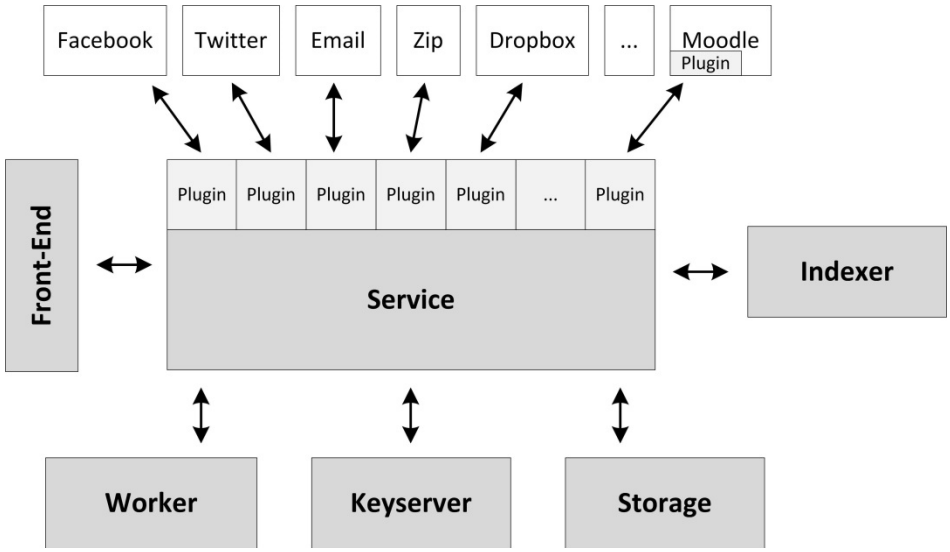

Abb. 1: Architektur von Themis

Der Sicherungsvorgang (Backup-Job) stellt die Kernentität des Systems dar. Ein Backup-Job besteht aus der sequentiellen Abarbeitung einer Datenquelle, mehreren Aktionen und einer Datensenke. Eine Datenquelle ist definiert durch einen Dienst von dem die zu sichernden Daten über eine wohldefinierte Schnittstelle heruntergeladen werden können, wie z.B. E-Mail, Twitter und Facebook. Aktionen sind Datenverarbeitungsschritte, wie beispielsweise die Extraktion und Anreicherung von Metadaten, deren Speicherung in einem schnell durchsuchbaren Index und das Erzeugen von Vorschaubildern. Der finale Schritt ist das Persistieren der Daten in einer Datensenke, wobei hier externe Dienste, wie z.B. Dropbox oder OneDrive, zur Verfügung stehen, als auch ein eigens konzipierter Themis Storage.

Front-End. Der Benutzer interagiert mit dem System über eine webbasierte Benutzeroberfläche. Sie bietet dem Benutzer eine einfach handhabbare und intuitiv zu bedienende Oberfläche für Registrierung und Anmeldung, Konto- und Profilverwaltung, Erstellung und Verwaltung von Sicherungsvorgängen, Authentifizierung externer Datenquellen und -senken, Suchen, Visualisierung und Aufbereitung der gesicherte Daten sowie darüberhinausgehende Themis Dienstleistungen wie die Möglichkeit, Daten zu vererben bzw. mit anderen Benutzern zu teilen.

Service. Die Themis Service Komponente steuert sämtliche Dienste im System, ist verantwortlich für die Verwaltung der Daten, Benutzer und Sicherungsvorgänge und stellt eine REST-API zur Interaktion bereit. Sie konfiguriert und verbindet Datenquellen, Datensenken und Aktionen zu Backup-Jobs. Diese werden über einen Plugin-Mechanismus bereitgestellt und können nach Belieben erweitert werden.

Worker. Worker sind die eigentlichen Instanzen, die für die Ausführung von Backup-Jobs verantwortlich sind. Sie können verteilt werden, um die parallele Ausführung von ressourcenintensiven Sicherungsvorgängen zu ermöglichen.

Storage. Der Themis Storage implementiert einen eigenen Dienst zur sicheren Speicherung von Daten und steht für Anwender der Themis Plattform mit einem vorgegebenen Quota zur Verfügung. Daten werden verschlüsselt und in einer anonymisierten Form abgelegt, die von außen betrachtet, keine Rückschlüsse über den Benutzer zulässt. Die Anbindung als Datensenke wird über ein Themis Storage Plugin bereitgestellt.

Indexer. Volltext- und Metadatensuche wird über eine Indizierung der Daten realisiert. Zur Aufbereitung der Suchergebnisse werden während eines Sicherungsvorganges die Daten mithilfe des Indexing Plugins analysiert. Dies beinhält Extraktion von Personen-, Orts- und Geodaten, Attribute zur Workflowausführung, sowie eine Analyse der Metadaten und Volltextextraktion mittels Apache Tika. Zur späteren Aufbereitung der Suchergebnisse werden die extrahierten Fragmente zur eigentlichen Indizierung an den Themis Indexer Dienst übergeben und anhand der definierten Sharing-Policy verteilt. Dieser Dienst ist einer der Kernstücke der Themis Sicherheitsarchitektur. Basierend auf Truecrypt für die Verschlüsselung und Elasticsearch für die Indizierung, wird hier für jeden Benutzer eine isolierte Instanz innerhalb einer verschlüsselten Partition gestartet.

Keyserver. Ein Hauptaugenmerk von Themis liegt in der sicheren Verwahrung der Daten. Hierbei soll der Benutzer jederzeit „Herr" über seine Daten sein. Ein unberechtigter Zugriff durch Dritte, oder durch den Betreiber selbst, soll unterbunden werden. Um dies zu gewährleisten wurden in der Themis-Plattform mehrere Sicherheitsmechanismen, basierend auf dem Keyserver Dienst, konzipiert und implementiert. Dieser erledigt systemweite Autorisierung von Datenzugriffen und Transaktionen und kümmert sich um das sichere

Speichern von Authentifizierungsinformationen. Alle Zugriffe auf diese Informationen werden validiert und protokolliert. Außerdem bietet der Keyserver das Abholen zeitlich begrenzter Security Tokens für das Ausführen von Backup-Jobs. Sensible Information werden durch eine wohlüberlegte Sicherheitsarchitektur gespeichert und sind für den Themis Systembetreiber zu keiner Zeit zugänglich.

Die Themis Kernmodule werden als Open Source Software unter der GNU General Public License oder vergleichbarer Lizenzen veröffentlicht. Der Source Code befindet sich auf https://github.com/backmeup/.

4 Security by Design

Der Ausgangspunkt der Themis Architektur liegt in der Ausarbeitung eines Security Konzeptes, welches die zu schützenden Werte, Bedrohungen und Sicherheitsziele beinhält. Anhand der nachfolgenden Anwendungsfälle *Backup erstellen, Backup ausführen und Backup Sharing* wollen wir einen detaillierteren Einblick in die Architektur der Themis Plattform bieten. Hierbei soll vor allem auf die Zusammenarbeit der Themis Subsysteme, einzelne technologische Architekturaspekte und sicherheitsrelevante Anforderungen und Lösungen eingegangen werden.

4.1 Backup erstellen

Beim Anlegen eines Backup-Jobs wird der Benutzer in der Weboberfläche schrittweise durch die verschiedenen Konfigurationsmöglichkeiten geführt. Zuerst wird eine Datenquelle gewählt. Zu jeder Datenquelle und Datensenke gehört ein Satz mit Authentifizierungsdaten. Dieser wird entweder durch OAuth-Authentifizierung (z.B. bei Facebook und Dropbox) erzeugt oder der Benutzer gibt die entsprechenden Daten (z.B. Benutzername und Passwort) direkt bekannt. Eine erfolgreiche Authentifizierung kann zur späteren Wiederverwendung abgespeichert werden. Nach der Auswahl der Datenquelle werden eine Datensenke zur Ablage der zu sichernden Daten sowie Aktionen für den Backup-Job ausgewählt. Die Konfiguration erfolgt dabei gleich dem oben beschriebenen Schema. Bei den Datensenken unterscheidet die Themis-Plattform aktuell zwei Arten von Speicherzielen: 1) Einfache Datensenken, welche nur die Ablage der gesicherten Daten auf einem Speicherplatz erlauben (z.B. Dropbox). Hier kann der Benutzer später über die vorhandenen Schnittstellen (Webinterface, Client-Apps) auf das Backup zugreifen. 2) Erweiterte Datensenken, welche neben der Ablage der gesicherten Daten auch einen Online-Zugriff auf diese erlauben (wie etwa das Themis Storage Plugin). Dies ermöglicht es dem Benutzer bei Abfragen im Themis-Index direkt das Suchergebnis aufzurufen und in den gesicherten Daten zu browsen. Aktionen bieten dem Benutzer die Möglichkeit, die gesicherten Daten während des Backups aufzubereiten, z.B. in Langzeitarchivformate zu konvertieren oder in besonderen Kompositionen anzuzeigen.

Für die oben genannten Bereiche bietet das Service eine flexible und modulare REST-Schnittstelle. Diese ermöglicht es Clients/dem Front-End, die vorhandenen Plugins für Datenquellen, -senken, Aktionen und deren Konfigurationsmöglichkeiten abzufragen. Auch können Authentifizierungs- und Konfigurationsprofile bei weiteren Backup-Jobs wiederverwendet werden. Die vom Benutzer angegebenen oder durch OAuth erhaltenen Authentifizierungsdaten werden im Keyserver durch einen symmetrischen Schlüssel (Datenquellen/-senken-Schlüssel) verschlüsselt abgelegt. Dieser Schlüssel und weitere benutzerspezifische Daten, z.B. das Benutzerprofil und die Public/Private-Keys für die

Datenverschlüsselung, sind durch einen weiteren symmetrischen Account-Key verschlüsselt. Der Account-Key selbst ist nur durch den Benutzer über Benutzername und Kennwort entschlüsselbar. Um auch diese beiden Informationen während der Benutzerinteraktion mit der Weboberfläche und den REST-Interfaces weitgehend zu verbergen, wird dem Benutzer nach erfolgreichem Login, ein einmaliges und zeitlich begrenztes UI-Login-Token zugewiesen. Wird nun vom Benutzer ein neuer Backup-Job angelegt, so werden mittels des eingeloggten Benutzers bzw. seines Account-Keys die erforderlichen Datenquellen-Schlüssel und der Public-Key des Benutzers entschlüsselt und innerhalb eines neuen Datensatzes als verschlüsselte Kopie gespeichert. Hierfür wird ein Onetime-Token generiert, das dem Backup-Job beigefügt und später vom Themis Worker bei der Ausführung verwendet wird.

4.2 Backup ausführen

Der Themis Worker übernimmt das eigentliche Abarbeiten der Backup-Jobs. Dabei werden folgende Schritte ausgeführt: 1) die Authentifizierung und der Download der Daten von der gewählten Datenquelle, 2) das Ausführen der gewählten Aktionen auf den heruntergeladenen Daten und 3) die Authentifizierung bei der Datensenke und das Sichern/Hinaufladen der Daten zu dieser. Treten während der Ausführung Fehler auf (z.B. eine Datei konnte nicht heruntergeladen werden) kann, wenn gewünscht, die Ausführung automatisch wiederholt werden.

Alle Teilaufgaben eines Backup-Jobs werden von spezialisierten und modularen Plugins erledigt, welche mittels OSGI geladen werden. Das Herunterladen von Datenquellen-Plugins, wie z.B. für Facebook und E-Mail Accounts, das Transformieren und Indizieren der Daten durch spezielle Aktionen-Plugins, z.B. Indizierung, Thumbnailgenerierung, und das Hinaufladen der Daten von Datensenken-Plugins, wie z.B. Themis Storage oder Dropbox. Dieser Ansatz ermöglicht eine einfache und flexible Erweiterung der Themis-Plattform um zusätzliche Datenquellen/-senken und Funktionalitäten.

Durch die Verwendung von Onetime-Tokens bei Backup-Jobs ist es nicht notwendig, dass der Benutzer zur Zeit der Abarbeitung seines Backup-Jobs eingeloggt ist. Gleichzeitig sind aber, im Gegensatz zu ähnlichen Konzepten bzw. Produkten, seine Authentifizierungsdaten nicht direkt und abrufbar im Backup-Job gespeichert, wo sie durch Angreifer und den Betreiber ausgelesen werden könnten. Die generierten Onetime-Tokens können nur einmalig und innerhalb eines gewissen Zeitfensters bezogen und verwendet werden. Wird ein Backup-Job periodisch ausgeführt, so kann beim Einlösen des Onetime-Tokens bereits ein weiteres für die nächste geplante Ausführung mitangefordert werden.

Die heruntergeladenen Daten werden im Themis Storage symmetrisch verschlüsselt gespeichert. Der zugehörige Schlüssel (Backupdaten-Schlüssel) wird nun mit dem Public-Key des Benutzers asymmetrisch verschlüsselt. Dies ähnelt dem bei PGP eingesetzten Verschlüsselungskonzept: PGP verschlüsselt den Datenstrom symmetrisch mit einem Schlüssel S. Dieser Schlüssel S wird wiederum mit dem öffentlichen Schlüssel PUB_A des Benutzers A verschlüsselt und in den Header der (verschlüsselten) Datei D geschrieben. Wird nun ein weiterer zugriffsberechtigter Benutzer B vom Benutzer A hinzugefügt, so wird zuerst S mit dem privaten Schlüssel $PRIV_A$ von A entschlüsselt und danach wieder mit dem öffentlichen Schlüssel PUB_B von B verschlüsselt. Der so generierte Eintrag wird wieder in den Header geschrieben, was allerdings ein Kopieren – „Nachrücken“ – der gesamten Datei (zu D’) erfordert und bei großen Dateien zu unnötigen I/O Operationen führt (siehe Abb. 2). Grundsätzlich gestattet PGP konzeptuell die Methodik des nachträglichen Hinzufügens eines

Empfängers, allerdings kommt diese selten zum Einsatz und ist daher auch nicht in gängigen und geeigneten PGP-Bibliotheken implementiert. Themis realisiert einem dem PGP ähnlichen, aber leicht abgewandelten Ansatz, bei dem der asymmetrisch verschlüsselte Schlüssel S in einer eigenen separaten Schlüsseldatei D_S abgelegt wird. Wird hier ein neuer Benutzer hinzugefügt, so funktioniert dies analog des PGP Vorgehens, aber der neue Eintrag wird nicht im Header der Datei D, sondern in der separaten Schlüsseldatei D_S eingefügt (siehe Abb. 2). Durch die Änderung in der relativ kleinen Schlüsseldatei umgeht man die Problematik, den gesamten verschlüsselten Datenstrom nachrücken zu müssen. Das implementierte Konzept ermöglicht sicheres und nachträgliches Hinzufügen weiterer Benutzer und somit deren Zugriff auf die gesicherten Daten. Ein Datenexport aus Themis erfolgt im Standard PGP Format.

Alle genannten Schlüssel und Tokens, außer des symmetrischen Schlüssels in der Schlüsseldatei, sowie die bereits genannten Authentifizierungsdaten, Benutzerprofile etc. werden im Keyserver-Dienst gespeichert und sind bei Bedarf nur vom eingeloggten Benutzer über sein Login-Token oder vom Worker über das Onetime-Token des Backup-Jobs abrufbar. Durch die Public-Private- Key-Verschlüsselung der Daten und des Suchindex (siehe Anwendungsfall Backup Sharing) kann somit nicht auf bereits gesicherte Daten des Benutzers, ohne dessen Zustimmung, zugegriffen werden. Übrig bleibt einzig ein kurzes Zeitfenster während bzw. nach dem Download sowie während der Transformation der Daten auf der Betreiber-Plattform. Hier könnte ein potentieller Angreifer oder der Betreiber durch direkten Zugriff auf diesen Datenfluss persönliche Daten und Informationen einsehen. Durch geeignete Sicherheits-, Monitoring- und Auditmaßnahmen kann die Wahrscheinlichkeit eines solchen unentdeckten Zugriffs vermindert und eingeschränkt, aber aufgrund der Notwendigkeit unverschlüsselter Daten in diesem Moment, nicht verhindert werden.

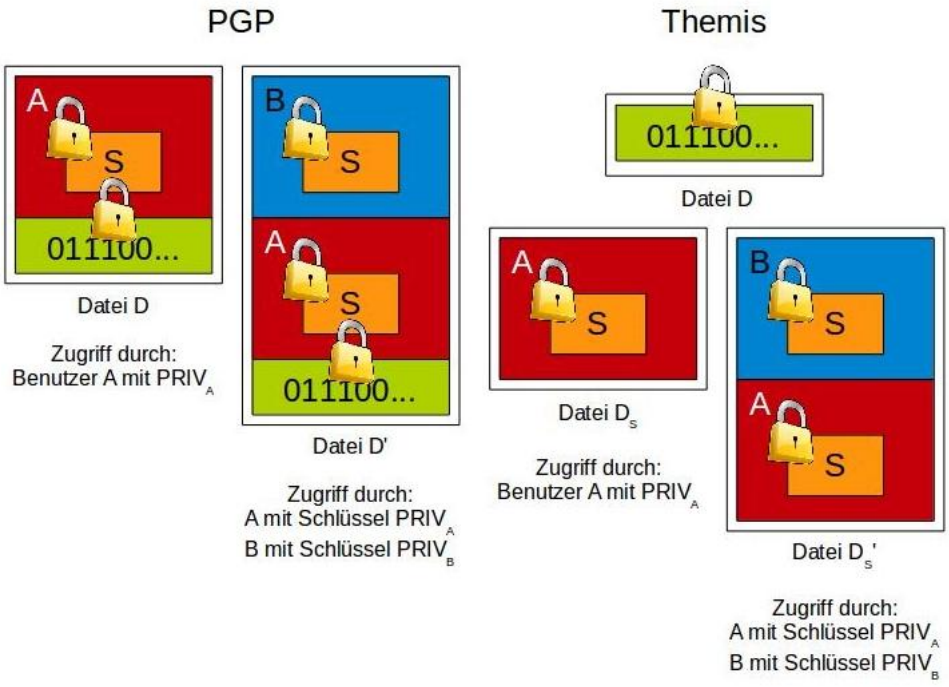

Abb. 2: PGP vs. Themis

4.3 Backup Sharing

„Backup Sharing“ bedeutet, bereits archivierte Daten anderen Benutzern zur Verfügung zu stellen und stellt eine jener Aktionen dar, die der Benutzer auf sein digitales Erbe auf der Themis Plattform setzen kann. Themis bietet somit, nicht nur im Fall des Vererbens, einen Raum, der es ermöglicht, Nutzungsrechte an digitalen Daten einfach und sicher an andere Benutzer zu übergeben. Über frei definierbare Sharing Policies kann die Granularität der zu teilenden Datensätze bestimmt werden. So ist es beispielsweise möglich, gesamte Backups, Suchergebnisse die einem bestimmten Begriff entsprechen, einzelne Dokumente oder frei durch den Benutzer zusammengestellte Datensammlungen, jetzt und auch zukünftig, mit einem anderen Benutzer zu teilen. Ein Handshake-Mechanismus garantiert die Bereitschaft des Gegenübers die Daten annehmen zu wollen. Daten eines anderen Benutzers weiter zu teilen wird hierbei technisch unterbunden. Realisiert wird dieser Anwendungsfall über den Themis Index-Dienst. Die digitale Nachlassverwaltung, also das Vererben im Ablebensfall, stellt einen Sonderfall dar, der aber Großteils über dieselben Mechanismen verwirklicht wird.

Da Themis Daten aus einer Vielzahl an Quellen sichert, ist eine einfache, schnelle und bedeutungsvolle Suche von zentraler Bedeutung. Der hierfür benötigte Index enthält hochsensitive persönliche Information, denn er stellt die gesicherten Daten bzw. die gesamte Historie des digitalen Lebens in komprimierter und abfragbarer Form dar. Die Innovation im Index liegt neben der Volltext- und Metadatensuche in der Extraktion und Aufbereitung von raum-, zeit- und personenbezogenen Informationen. Kommunikation, Bilder, Termine und Dokumente, all diese digitalen Entitäten entstehen in einem Kontext. Dieser Raum-/ Zeit-/Personenbezug gibt den Kontext der digitalen Daten im Umfeld wieder und bietet so einen innovativen Mehrwert in der Auffindbarkeit und Aufbereitung der Suchergebnisse für den Benutzer der Themis Plattform. Themis gibt klare Sicherheitsanforderungen hinsichtlich der strikten Trennung von Userdaten vor. Daten befinden sich verschlüsselt am Filesystem sowie im Index, und werden nicht in Datenbanken vorgehalten. Es muss sichergestellt werden, dass nicht nur der Zugriff auf das digitale Erbe, sondern selbst die bloße Existenz eines Datensatzes vor anderen Benutzern sowie vor dem Betreiber des Systems, verborgen ist.

Als Suchservertechnologie wird das auf Apache Lucene basierende Open-Source Tool Elasticsearch verwendet, welches auf Skalierbarkeit ausgelegt ist [KuRo13] und eine Integration distanzbezogener Geokoordinatenabfragen mitbringt. Eine projektinterne Evaluierung hat ergeben, dass die zuvor genannten Anforderungen aber nur teilweise über die Elasticsearch Multi-Index API realisierbar sind. Während mit der PGP-ähnlichen Dateiverschlüsselung in Themis eine Duplizierung der Backup-Dateien selbst im Anwendungsfall des Backup-Sharing vermeidbar ist, ist es notwendig, die jeweiligen Index-Fragmente zu duplizieren und einen isolierten Index pro Benutzer zu verwalten. Das Index-Core Modul in Themis realisiert das Index-Pro-User Modell auf Basis von Elasticsearch (Version 1.4.0) und Truecrypt (Version 7.1a). Der Einsatz von TrueCrypt wird auf Basis des Open Crypto Audit Project (OCAP)[2] als bedenkenlos eingestuft. Für jeden Benutzer der Themis-Plattform wird ein verschlüsselter TrueCrypt Container erzeugt, welcher bei Login des Benutzers eingehängt wird. Die Cluster-Konfiguration der Elasticsearch Instanz wird für den Benutzer dynamisch erzeugt und eine isolierte Suchinstanz innerhalb des verschlüsselten

[2] Endbericht des Open Crypto Audit Project (OCAP) von TrueCrypt Version 7.1a https://opencryptoaudit.org/reports/TrueCrypt_Phase_II_NCC_OCAP_final.pdf

Containers hochgefahren. Hierbei ist eine Querkommunikation zwischen aktiv laufenden Instanzen verschiedener Benutzer strikt unterbunden.

Nachdem die Suchinstanz nur während der aktiven Zeit des Benutzers im System zur Verfügung steht, Backup-Jobs aber asynchron und einem frei definierbarem Sicherungsintervall folgend, ausgeführt werden können, ist es notwendig, die eigentliche Aufnahme der Datensätze in den Index von der Indexfragmenterzeugung zu entkoppeln. Indexfragmente sind aufbereitete Datenergebnisse aus dem Indexing Plugin, welches innerhalb der Job-Abarbeitung im Worker zum Einsatz kommt. Sie werden in serialisierter Form, gesteuert durch die definierten Sharing Policies, in den jeweiligen Drop-Off Bereichen des Benutzers sowie seiner Sharingpartner physisch persistiert (siehe Abbildung 3) Drop-Off Bereiche sind mit dem Public-Key des jeweiligen Benutzers verschlüsselt und erlauben somit dem Themis Worker Daten jederzeit sicher abzulegen. Betritt ein User nun das System, werden anstehende Import- und Löschoperationen auf dessen Inhalt nachgezogen und das physische Index-Fragment in den gemounteten TrueCrypt Container des Benutzers verschoben, wo es langfristig persistiert wird, um ein nachträgliches Teilen von Inhalten zu ermöglichen. Anstehende Updateoperationen auf den Index Content sind in Form von Taskbeschreibungen (wie z.B. toImport) und Fragment UUID in der Datenbank hinterlegt. Somit sind rasche Auswertungen auf Einhaltung der zugrundeliegenden Sharing-Policies auf Datenbankebene möglich.

Die Summe an Operationen stellt zu jeder Zeit einen konsistenten Status des Systems dar ohne jedoch sensible Informationen vorhalten zu müssen. Diese asynchrone Entkopplung und dynamische Instanzierung des Suchproviders, bei gleichzeitiger Nutzung der Skalierbarkeit und Effizienz des Elasticsearch Frameworks, stellt eine innovative Lösung und wesentlichen Bestandteil der Themis Sicherheitsarchitektur dar.

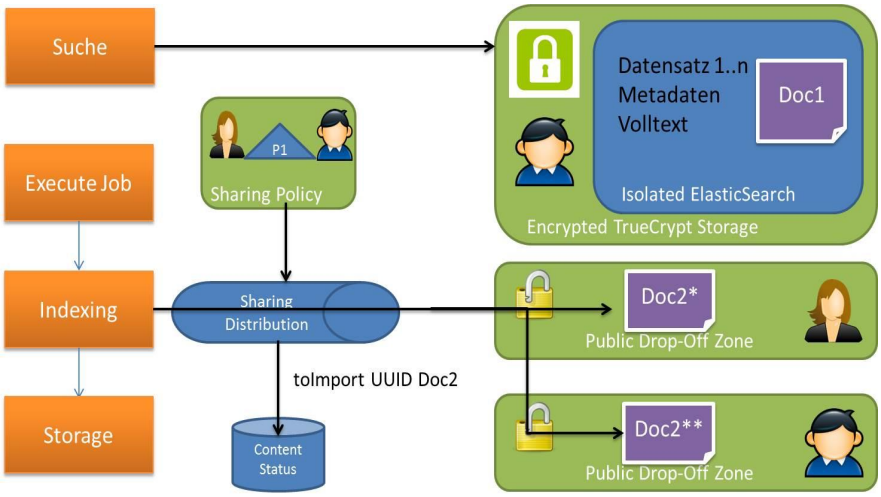

Abb. 3: Realisierung des Anwendungsfalls Backup Sharing

5 Ausblick

Themis ist ein automatisierter Backupdienst zur Sicherung, Verwaltung und Vererbung von persönlichen Daten. Möglichkeiten zum Stilllegen eines nicht länger benötigten Quellenaccounts oder zur Löschung der darin gespeicherten Daten werden auf Grund mangelnder Schnittstellen und restriktiver Nutzungsbedingungen der angebundenen Provider nicht angeboten. Der Aufwand neue Dienste und Plattformen in Themis zu integrieren ist

nicht unerheblich. Stabilere APIs und Standards wären wünschenswert. Weiterführende Initiativen werden sich aus diesem Grund auf die Bereiche der europaweiten Standardisierung von Schnittstellen zur Langzeitarchivierung, insbesondere der vollständigen Abrufbarkeit an Benutzerdaten im Zuge einer Accountstilllegung, einsetzen.

In der derzeitigen Ausbaustufe des Themis Systems vertraut ein Benutzer darauf, dass der Betreiber dieser Plattform weder beobachtend, noch manipulierend in das System eingreift. Themis kann selbst gehosted und betrieben werden. In der finalen Phase des Projektes wird es darüber hinaus möglich sein, private Instanzen von sensiblen Teilsystemen, wie beispielsweise den Worker oder den Storage Dienst, als private Cloud-Instanzen zu mieten und diese somit nicht mit anderen Benutzern zu teilen, oder diese selbständig unter eigener Kontrolle zu betreiben und in die Anwendung einzubinden.

Die Endphase des Projektes beschäftigt sich mit der Forschungsfrage der semantischen Interpretation der archivierten Daten und Verknüpfung von Entitäten über verschiedene Datenquellen eines Benutzers hinweg. So soll es beispielsweise möglich sein, Smartphone Backups mit SMS, Kontakten, Kalendereinträgen und Lesezeichen mit archivierten Dokumenten aus der Dropbox sowie Fotos von Desktop-Backups über Zeit-, Personen- und Ortsfilter zu durchsuchen und diese für den Benutzer in einer Timeline Aufbereitung zur Navigation zur Verfügung zu stellen.

Darüber hinaus gilt die exakte Skalierbarkeit des Systems zu testen und den Ressourcenverbrauch des Index-Per-User Modells, wie durch den Elasticsearch-in-TrueCrypt Ansatz realisiert, zu evaluieren.

5.1 Danksagung

Das Projekt Themis wird von der Österreichischen Forschungsförderungsgesellschaft (FFG) im Rahmen der Programmlinie COIN Kooperation und Netzwerke mit Mitteln des Bundesministeriums für Verkehr, Innovation und Technologie und des Bundesministeriums für Wirtschaft, Familie und Jugend gefördert. Das präsentierte System basiert auf der Arbeit der Projektpartner X-Net Services GmbH, AIT Austrian Institute of Technology GmbH, FH OÖ Forschungs & Entwicklungs GmbH, Johannes Kepler Universität Linz – Institut für Datenverarbeitung in den Sozial- und Wirtschaftswissenschaften, GTN – global training network GmbH, MIRACLE Information Systems GmbH, Dietmar Gombotz S3 – Software, Systems, Services Dietmar Gombotz. Getroffene Aussagen, Schlussfolgerungen und Meinungen in diesem Paper, sind jene der Autoren und repräsentieren nicht zwangsläufig die Ansichten des Fördergebers.

Literatur

[GaRe12] J. Gantz, D. Reinsel: The Digital Universe: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. In: IDC iView: IDC Analyze the Future (2012) 1-16.

[OSHT12] W. Odom, A. Sellen, R. Harper, E. Thereska: Lost in translation: understanding the possession of digital things in the cloud. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2012) 781-790.

[McNS09] F. McCown, M. L. Nelson, H. van de Sompel: Everyone is a curator: Human-Assisted Preservation for ORE Aggregations. In: CoRR (2009).

[HLT+13] R. Harper, S. Lindley, E. Thereska, R. Banks, P. Gosset, G. Smyth, W. Odom, E. Whitworth: What is a file? In: Proceedings of the Conference on Computer Supported Cooperative Work (2013) 1125-1136.

[MaSh14] C. C. Marshall, F. M. Shipman: An argument for archiving Facebook as a heterogeneous personal store. In: Proceedings of Joint Conference on Digital Libraries (2014) 11-20

[Chau14] P Chauving: The race to archive Twitpic before 800 million pictures vanish. In: The Globe and Mail (2014).

[DJRW10] K. Dean, J. L. John, I. Rowlands, P. Wiliams: Digital Lives. Personal Archives for the 21st Century: An Initial Synthesis. (2010)

[LMB+13] S. Lindley, C. C. Marshall, R. Banks, A. Sellen, T. Regan: Rethinking the Web As a Personal Archive. In: Proceedings of the International Conference on World Wide Web (2013) 749-760.

[MaSh11] C. C. Marshall, F. M. Shipman: The ownership and reuse of visual media. In: Proceedings of Joint Conference on Digital Libraries (2011) 157-166.

[GOFF13] R. Gulotta, W. Odom, J. Forlizzi, H. Faste: Digital Artifacts as Legacy: Exploring the Lifespan and Value of Digital Data. In: Proceedings of the SIGCHI Converence on Human Factors in Computing Systems (2013) 1813-1822.

[StPR11] S. Strodl, P. Petrov, A. Rauber: Research on Digital Preservation within projects co-funded by the European Union in the ICT programme. (2011)

[SCM+13] S. Schlarb, P. Cliff, P. May, W. Palmer, M. Hahn, R. Huber-Moerk, A. Schindler, R. Schmidt, J. van der Knijff: Quality assured image file format migration in large digital object repositories. In: Proceedings of the International Conference on Digital Preservation (2013).

[GSPR10] M. Greifeneder, S. Strodl, P. Petrov, A. Rauber: HOPPLA - Archiving System for Small Institutions. In: ERCIM News (2010).

[KuRo13] R. Kuc, M. Rogozinski. ElasticSearch server. In: Packt Publishing Ltd (2013).