Themis - Conserve Your Digital Life

Sebastian Pimminger [a], Johann Heinzelreiter [a], Werner Kurschl [a], Andrew Lindley [b]

[a] FH OÖ Forschungs & Entwicklungs GmbH, Softwarepark 11, A-4232 Hagenberg, AUSTRIA
[b] AIT – Austrian Institute of Technology GmbH, Donau-City-Strasse 1, A-1220 Wien, AUSTRIA

ABSTRACT:
Today the digital heritage of millions of users is in the hands of a few providers of web services and social networks. These can disappear overnight or no longer make their services available. Sometimes it can have serious consequences if a service is not reachable for days or even hours .In this paper the authors describe the Themis platform. Themis is dedicated to this personal digital identity of a user and allows cross-platform backup of distributed online platforms, services, mobile and desktop sources. The platform also addresses the legally regulated transmission of data between users and from one generation to the next. Furthermore, objectives of long-term archiving, data characterization and full-text searchability with space/time affinity as well as scalability are taken into account. All of this adds up within the Themis platform to provide users with an added value upon their data and is done under strict safety and privacy requirements to prevent against illegal insights.

1 INTRODUCTION

According to a study by IDC [1] the digital universe, i.e. the full set of data stored worldwide, will double every two years, reaching 40,000 exabytes in 2020. The study also states that already in 2012 68% of the data is created and consumed by customers. At the same time, it can be observed that cloud computing is growing at a rapid pace. People are interacting with social media, they are storing camera phone images and videos on cloud platforms, and they are entrusting more and more of their private documents to cloud storage providers [6]. In the aforementioned study IDC estimates that by 2020 nearly 40% of data in the digital universe will be stored or processed by cloud providers.

The rapidly changing situation in our digital environment raises and number of questions:

- The data of millions of users – their digital heritage – is spread over a large number of online services. Many users have lost overview over the service providers they consume or are even not aware of the fact that their data is stored in the cloud.
- Consequently, for many users it is difficult if not impossible to find the dedicated storage location of their documents. This problem is even reinforced by the fact that users collect data over many years and decades [7].
- Data stored in the cloud may be destroyed accidentally or may become inaccessible over time or will even be deleted when the user forgets to pay for the service.
- For the majority of service providers it is totally unclear how to treat customer data if users decide to suspend their service subscription [2].
- Most often than not data in the cloud is totally unprotected. Recent reporting about internet surveillance by American secret services have increased the willingness to rely on cryptography to protect the users' data.
- Social media networks and other online services have emerged only over the last ten years. Nevertheless, there are more and more reports that relatives of deceased persons have problems to get access to the deceased's online services and data. It may be tedious to find out which services the deceased person used, even the deletion of an online account may be an unsolvable problem.

There are many tools and systems that address the requirements mentioned above. But all of them are restricted to only a few of these aspects. A lot of tools deal primarily with backup, like ArchiveFacebook[1], Gmail Backup Tool[2], or Backupify[3]. Many of them are restricted to specific online services and are not extendable. None of them is able to build an index on the collected data which would create additional value of helping users to find their way in the maze of data. Other services focus on encryption, e. g. Wuala[4], but lack automation of the backup process. The mobile app Timehop[5] for iOS and Android focuses on an integrated user experience and lets you explore pictures and postings of selected social media sources and the phone's camera roll as they happened, day by day in the past, but lacks harvesting and preserving this data under the user's control.

When it comes to securely sharing and inheriting data we enter totally new territory. Up to now, only very few projects and tools have touched this area [8]. The cloud-based password administration tool PasswordBox[6] has a legacy extension that allows the inheritance of access data to online accounts to one's heirs. VitalLock[7] provides encrypted online storage and enables users to send vital information in case of emergency or death.

As can easily be seen, all these tools have a very limited scope and are restricted to backup, encryption of data, or the exchange of simple messages between the testator and the heirs. We strongly believe that there is the need for a system that fulfills more or less all requirements specified above. Such a tool could help people to regain sovereignty of their digital heritage.

2 OBJECTIVES

The technical platform of Themis derives from is the BackMeUp project, which was developed from 2011 to 2012, supported by FFG funding. This initial proof-of-concept implementation focused on demonstrating the idea of personal web archiving by providing plugins for selected data sources – i.e. pulling in social media streams such as pictures, comments, friend's lists from Facebook or Twitter – and data sinks as to Dropbox or zip containers to which the harvested information was written. The goal was to archive personal information from online sources within the user's sphere of influence for preservation purposes. The required workflow was delivered through the BackMeUp environment. This included a secure container for storing highly sensitive data such as plugin authentication information or user credentials for scheduled recurring backups, a simple workflow execution environment as well as a graphical user interface which was able to deliver full-text searchability based on Elasticsearch and indexed metadata derived from Apache Tika analysis [5]. However as all information was meant to be transient, only metadata records were kept on the portal which led to poorly integrated search results and a bad user experience. The list of plugins finally was extended with one for email (POP3 and IMAP) archiving as well as a plugin for the online-learning platform Moodle so that students of the Johannes Kepler University could keep a personal record of their attended courses.

Themis supersedes the idea of BackMeUp by raising the objective to provide an integrated service for dealing with private digital heritage. This includes the areas of generating generation backups in terms of defining subsets of digital content of multiple data sources - such as digital picture collections, documents that personally matter, including the mobile phone as data source. This data is harvested, encrypted, preserved and only accessible to the Themis user himself. Using state of the art encryption mechanisms throughout the entire architecture pre-

[1] https://addons.mozilla.org/en-US/firefox/addon/archivefacebook (last visited Jan. 2015)

[2] http://www.en.gmailbackup.org (last visited Jan. 2015)

[3] https://www.backupify.com (last visited Jan. 2015)

[4] www.wuala.com (last visited Jan. 2015)

[5] http://timehop.com/ (last visited Jan 2015)

[6] https://www.passwordbox.com (last visited Jan. 2015)

[7] http://vitallock.com (last visited Jan. 2015)

vents a Themis system provider to access actual unencrypted data at any time except when the user grants private key privileges. One of the benefits of our underlying encryption architecture is that the Themis platform is able to offer the possibility of secure content sharing. Individual items, individually composed data sets or even entire backups may be securely shared with either other platform users or, for the event of death, deposited for handover as generation backup to ancestors within a notary act.

Themis is able to offer a full-text search across your archived data of online services, files from your desktop and mobile device. Bringing in additional records such as calendar entries, short message chat histories or browser bookmarks are seen as key objective towards success, as keeping data accessible from these rapidly emerging but frequently changing environments is a preservation challenge which generates added value for the private sector and on the other hand leverages the possibilities of the platform in offering an integrated experience in the areas of contextual search. Themis is able to deliver meaningful search results the user cares about, for example by finding data and also linked data items, by relating items to geo- and temporal metadata such as a specific places or an event in time. Themis should be able to respond to queries such as give me all files, SMS and emails edited on the 29th of January 2015, show me pictures taken near Vienna.

Beyond this, advanced mechanisms of exploring mobile content, for example providing emulation environments for replaying access to mobile records as apps and the possibility of including home NAS systems as data sinks are being evaluated.

An analysis of competitors in the field showed interesting offerings, but all fell short in providing an integrated experience for the digital heritage of private information. Available platforms either try to act as reliable, trustworthy and secure authorities, which hand over information such as passwords or user credentials for online-services to beneficiaries[8], provide data forensics as extracting relevant data such as photos[9], files or user-profiles or act as safety deposit box within the cloud to store highly sensible data in a secure and encrypted manner[10]. Prominent data hosting services and public cloud offerings such as Google Drive, Dropbox, etc. are very popular for storing information mainly because of immediate access, while the underlying data regulations are vague. Themis however places the objective to implement a highly secure and trustworthy service for digital content, governed by Austrian law. Even though legal bodies like the notary's office for handling the process of private digital heritage are included as an actor within the platform – the main goal is to provide a highly attractive, effective and useful service during the entire digital lifetime of a user and beyond. A highly trustworthy and secured service combined with a simple way of backing up information from various data sources, providing the possibility of finding information based on full-text-search or contextual data such as places or time spans, together with the ability of sharing information with other system users and downloading all information at any time if necessary makes Themis a highly attractive service offering.

3 METHOD

The basic method used to develop Themis is a UX process, as described in [4], which consists of the four activities i) analyze, ii) design, iii) implement/prototype, iv) evaluate which are carried out iteratively. During the project, we iterated through these actions and often moved back to previous activities. In the analyze activity, we tried to understand user requirements and needs. In the design activity, we created interaction design concepts. In the implement/prototype activity, we realized design alternatives and in the final activity of the process lifecycle, we verified and refined our interaction design.

We started our UX process with some basic assumptions, which where aligned with a number of potential customers, so that our starting point is set well beyond a classical backup tool. First,

[8] www.securesafe.com (latest visit Jan. 2015)

[9] www.semno.de (latest visit Jan. 2015)

[10] legacylocker.com (latest visit Jan. 2015)

even if we build a system for digital heritage, it is not the most important intention of the ordinary user, to think about situations occurring far in the future. Thinking about inheritance starts usually quite late, when users have established their family and built up some assets. Second, to bring some early benefits for users, we assumed that discovery of digital assets and sharing content with friends and family members might be of equal value, if we provide a secure and easy to use service, which goes well beyond classical photo and document sharing services.

Third, storing data must be done in such a way, that users aren`t affected by running their own hardware, running out of storage space, protecting the storage facility from physical damage or theft, and that the data is available 24 hours, 7 days per week, and 365 days per year. This led us immediately to a cloud computing approach. Fourth, conserving data for a long period of time is not only a technical challenge, but also a question of how we can ensure business continuity. Since companies may fail, we set the goal that the solution must be open-source, so that Themis may run on most of the typical datacenters in the world. Fifth, the initial intention of keeping the notary in the loop, remains valid, but it is a long-term investment, since the user-base of digital natives is steadily growing throughout all ages. This implies that necessary workflows should easily fit into the daily life of Themis users, but also comply with national regulatory issues of the inheritance law and the business of civil law notaries.

4 ARCHITECTURE

The overall architecture of the Themis system is shown in Figure 1. The prototype is structured in five major subsystems that are organized as services. These services are self-contained and exchange information and data through well-defined interfaces.

A Backup job is the main entity of the system. All contributing services provide functionality to compose, distribute and execute them and their sensible information in a secure way. Operators may be restricted as much as possible to access this sensible information Therefore, we provide for example special data stores (see Keyserver). Generally, Backup jobs are described as simplified pipelines containing one data source, multiple actions, and one data sink. A data source is defined by a service where data can be downloaded. Examples of these data sources include e-mail, Twitter or Facebook accounts. In next step of the pipeline, this data may be processed or transformed by actions. For example, it is possible to extract information of this data and store it in a fast searchable index or apply a state of the art encryption algorithm to the data. At the end, all data is uploaded to a data sink (e.g. Dropbox or a downloadable zip archive).

To extend the system for arbitrary data sources, data sinks and actions, the prototype uses a plugin-based approach. Developers can contribute their own plugins and extend the systems functionality. Besides, this brings simplified deployment and offers higher customization support as plugins may be added, updated, or removed on the fly even from remote locations over a network. For executing backup jobs, the prototype distributes its workload between multiple workers. In doing so, the system follows a master/worker model [3]. In the following, the services are discussed in more detail.

Front-End. The web front-end is the main entrance point for end-users. It features a simple and intuitive graphical user interface for managing backups, account and profile information. It provides access and search through backup data and allows defining sharing policies between users. Furthermore, the front-end does not only target the classical web user, but comes within a responsive design for mobile platforms and as native iOS and Android app.

Service. The Service is the core component of the system. It integrates all other services and offers a well-defined API. The Service is responsible for managing data and provides central infrastructure services such as user and job management. It acts as master for the Worker component and composes data sources, sinks and actions to backup jobs. The Service schedules these jobs, distributes them among the Workers, and gathers information about their execution such as logs and progress.

Keyserver. The Keyserver provides a secure storage and access layer for personal keys and authentication information such as pre signed backupjob execution tokens, which can only be accessed in a limited time slot. Furthermore, it is responsible for core security concerns of the whole ecosystem such as user authentication and authorization. All sensitive information is encrypted and even prevents the Themis operator from accessing it.

Indexer. A distinguishing feature of this platform is its capability of full-text searchability through backup data and its metadata. Especially in combination with location and time aware data, it can answer questions related to special events or places. This makes it a powerful feature in the context of the digital life. Since a user's index may contain sensible information, we strictly separate them in an index-per-user model. This requires isolated Elasticsearch indices and their dynamic activation and deactivation on the fly As well as secure content drop-off zones which get re-ingested private key access is granted by the user on system login.

Worker. The Worker component is responsible for executing backupjobs defined by the Service. It executes in very simple cycle: receive a job from the Service, process it, and receive the next job and so on. The worker also uses the common plugin infrastructure described above. Typically, a system consists of multiple Workers to spread the potentially long running up- and download processes, data analysis and encryption tasks among them. In future releases, it may be considered that users can register their own dedicated private Worker that is only capable of processing jobs from its owning user to emphasize data ownership and security concerns.

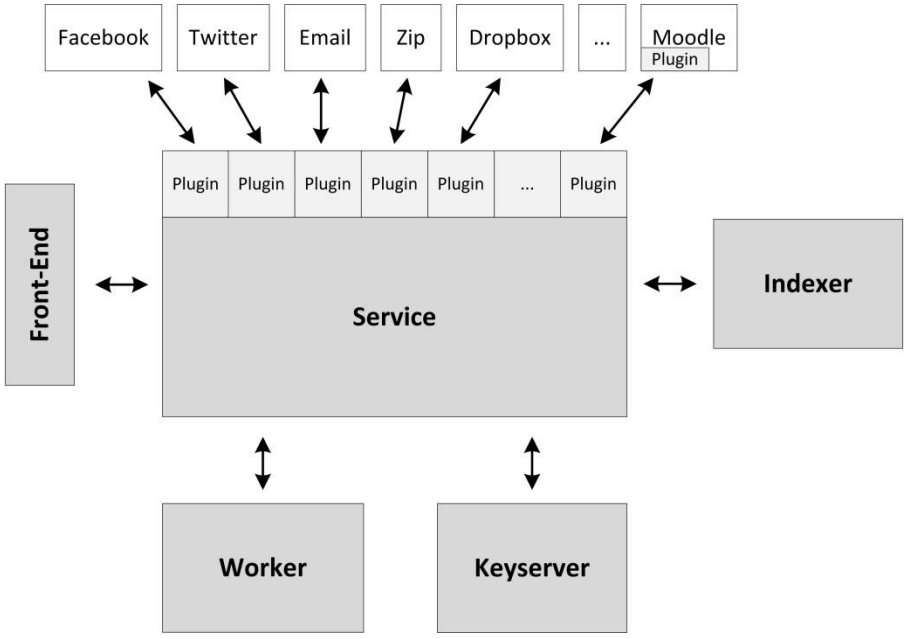

Figure 1. Architecture of the Themis prototype.

5 CURRENT STATE OF DEVELOPMENT

The services of the Themis system (Figure 1) are implemented as independent and exchangeable modules with a firm boundary. They interact in a RESTful way with each other. Although the different services may be implemented with different programming languages and technologies, we use Java 1.7 for the sake of simplicity to implement most parts of the system. Only for the front-end development, we leverage the Apache Cordova framework to build web and mobile UI applications from the same code base.

As mentioned above, the Indexer is one of the distinguishing features of the Themis ecosystem. To implement the index-per-user model, we use a combination of Elasticsearch (ES) and TrueCrypt[11] (TC). When accessing a user's index, the corresponding TC container is mounted, a user specific startup script is generated and an isolated index instance of ES starts up. The index of this specific user is now accessible and can be populated or queried. After a certain time of inactivity or explicit shutdown, the ES index is closed and persisted back to the TC container which is then unmounted. With this approach, scalability is certainly an issue when serving a large number of concurrent requests. Therefore, we are currently limited to the performance of the machine running the Indexer service. The individual layers have already been separated so that spreading the load to multiple index node servers will be addressed in future releases.

Up to now several source, sink, and action plugins have been implemented. The following plugins are currently available (source plugins are marked as (+), sinks as (-) and actions marked as (o); plugins printed in italic are currently under development): Email (+), Facebook (+), *Moodle* (+), *Twitter* (+), Dropbox (+/-), *Skydrive* (+/-), Zip (-), *Encryption* (o), Indexing (o), Thumbnail (o). The architecture of the plugin API has proven to be flexible enough to support a wide range of heterogeneous services and technologies. Currently the development of a *SFTP* (+) plugin is carried out which will be used as interface for mobile and desktop content synchronization within the workflow.

At typical backup workflow is shown in Figures 2 – X. For testing purpose, we use multiple accounts (e.g. Facebook, Gmx Email, Twitter) filled with realistic test data. For example, the Themis Facebook account is populated with multiple posts, photos with comments, events, and linked friends). Although we don't have reliable data yet, typical backup jobs with indexing enabled can be characterized as long running (depending on the download and upload size) and CPU intensive (due to full text and metadata extraction). Different plugins however impose diverging demands on the Worker and therefore determine how many jobs can be executed in parallel on a single machine.

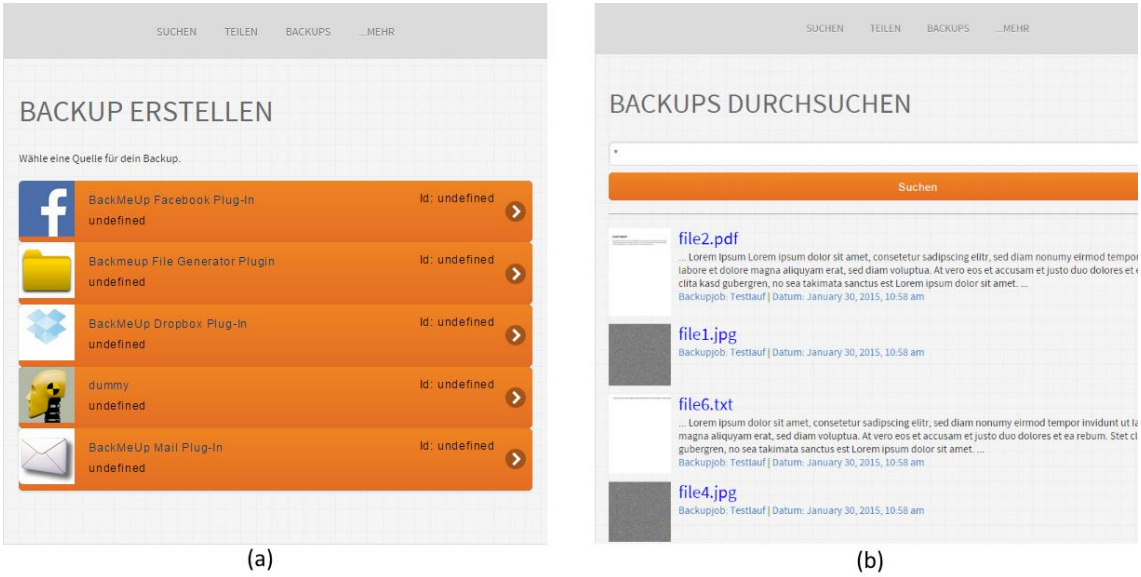

Figure 2. Screenshots of Themis front-end. Figure (a) shows an overview of all available data source plugins. Figure (b) shows the search results from the indexed backup data.

[11] http://truecrypt.sourceforge.net/ (last visited Jan. 2015)

Themis is freely available under GNU General Public License and can be downloaded from GitHub (https://github.com/backmeup). Additional information can be found at the project homepage http://www.backmeup.at

6 CONCLUSION

In this paper we have outlined the challenges and objectives underlying the Themis project and presented the current state of software development. Themis supports the secure preservation and management of personal digital assets from various data sources (e.g. Facebook, smartphones, cloud stores, mobile and desktop, etc.). While all preserved data is kept in an encrypted version accessible and downloadable anytime for the user Themis as a platform applies added value such as full-text search and metadata extraction on top. Themis applies a novel inheritance concept that allows to easily share specific data items such as pictures or documents with other people and in the worst case to the legal successor. Next, we want to evaluate this novel concept to gain a deeper understanding of usage patterns like public notary involvement, resource consumption, scalability and data fragmentation as well as key management issues. A long term analysis with UX metrics, as proposed by [4], will probably reveal further insights. Finally we will make use of the extracted metadata and data provided from mobile backups to enhance the platforms search experience in a spatial and temporal manner.

ACKNOWLEDGMENT

This research was supported by a grant from the Austrian Research Promotion Agency (FFG) under the COIN program. The partners of this project are Johannes Kepler University (Department of Data Processing in Social Sciences, Economics and Business), University of Applied Sciences Upper Austria (Software Engineering), Austrian Institute of Technology, XNet, gtn, Miracle Information Systems, and Dieter Gombotz S3. Any opinions, findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the research sponsors.

REFERENCES

[1] John Gantz and David Reinsel: *THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, 2012* http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

[2] Ernie Smith: *What we lose when an online startup fails*, last visited Jan. 2015, http://associationsnow.com/2014/10/twitpic-online-startup-fails/

[3] David. P. Anderson, *BOINC: A System for Public-Resource Computing and Storage*, in Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing, IEEE Computer Society, 2004, pp. 4–10.

[4] Rex Hartson, Pardha Pyla: *The UX Book - Process and Guidelines for Ensuring a Quality User Experience*, Morgan Kaufmann, 2012

[5] Markus Radtisch, Peter May, Asger A. Blekinge, Per Moldrup-Dalum: *SCAPE - Characterisation Technology, Release 1 and Release Report*, 2012, last visited Jan. 2015 http://www.scape-project.eu/wp-content/uploads/2012/03/SCAPE_D9.1_SB_v1.0.pdf

[6] William Odom, Abigail Sellen, Richard Harper, Eno 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, ACM, 2012, pp. 781-790.

[7] Sian Lindley, Catherine C. Marshall, Richard Banks, Abigail Sellen, Tim Regan: *Rethinking the Web As a Personal Archive*, in Proceedings of the International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2013, 749-760.

[8] Rebecca Gulotta, William Odom, Jodi Forlizzi, Haakon 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, ACM, 2013, pp. 1813-1822