Building an Academic Cloud for 500,000 Users

Experiences from the "sciebo" Project

Raimund Vogl
rvogl@uni-muenster.de

Holger Angenent
h.angenent@uni-muenster.de

Dominik Rudolph
d.rudolph@uni-muenster.de

Anne Thoring
a.thoring@uni-muenster.de

Christian Schild
schild@uni-muenster.de

Westfälische Wilhelms-Universität Münster
ZIV
Röntgenstraße 7-13
48149 Münster
Germany

Stefan Stieglitz
stefan.stieglitz@uni-due.de

Christian Meske
christian.meske@uni-due.de

University of Duisburg-Essen
Department of Computer Science and Applied Cognitive Science
Bismarckstrasse 81
47057 Duisburg
Germany

Abstract

Under the brand name "sciebo – the Campuscloud" (derived from "science box") a consortium of more than 20 research and applied science universities started a large scale cloud service for about 500,000 students and researchers in North Rhine-Westphalia, Germany's most populous state. Starting with the much anticipated data privacy compliant sync & share functionality, sciebo offers the potential to become a more general cloud platform for collaboration and research data management which will be actively pursued in upcoming scientific and infrastructural projects. This project report describes the formation of the venture, its targets and the technical and the legal solution as well as the current status and the next steps.

Keywords: Project report; cloud services; cloud storage; information infrastructure; IT infrastructure; university IT; collaboration; cooperative project

1. The Sync & Share NRW Project and Consortium

1.1. Project History

Since the start of Dropbox in 2007, cloud services have experienced a very fast diffusion. Dropbox alone has more than 300 million users (Dropbox et al. 2014). In Germany every fifth person uses cloud services (Eurostat 2014). The triumphal procession of these convenient, easy und mostly free to use services did not stop at the universities' doors, quite the contrary: About 80 percent of students and employees already use cloud services, as a study conducted in 2013 at three large German universities (RWTH Aachen, WWU Münster, University of Bonn) and with a total of 10,367 completed questionnaires (Meske et al. 2014, Stieglitz et al. 2014) shows. This is remarkable, because terms and conditions of those services force users to virtually surrender all data to the companies operating them. In some cases their use equals a violation of university regulations for the use of cloud services, national data protection laws or terms of articles of employment. But compliant solutions were hard to find.

In early 2012, a student initiative at the WWU Münster asked for a safe alternative where data is stored on the university's own servers. A representative user survey (Vogl et al., 2013) in the same year reassured the demand for such a service among both students and employees. Because of the high synergetic potential, the WWU's computing center (ZIV) decided against a solitary solution and in favour of a consortium approach with other research universities in the German state of North Rhine-Westfalia (NRW) to create an on-premise private sync & share cloud storage service. With potentially 500,000 employees and students as its users, it is to our knowledge one of the largest joint projects in the higher education IT.

With a positive review of the project funding proposal by the German Research Foundation (DFG) and 2.79 Million Euro in funding from the Ministry of Innovation, Science and Research NRW (MIWF), the project, led by the ZIV, was ready to lift off in early 2014. Preparatory work was finalized in mid 2014, building on continuous monitoring and evaluation of software solutions for on-premise sync & share (showing very good progress for the open source ownCloud project) and suitable hardware platforms (with a very promising proof of concept by IBM with its GPFS storage server (GSS) platform and ownCloud - demonstrating scaling for a 100,000 user per site setup) conducted since mid 2012. Furthermore, as mentioned above, a revised multi-university survey focusing on user expectations in cloud services and on demand for storage space and features to gauge the project objectives was conducted in 2013 and answered by over 10,000 employees and students at three major universities (Stieglitz et al., 2014).

A final evaluation of the available software solutions confirmed the conclusion that ownCloud was covering all features required for the project and was uniquely positioned as an open source solution, enabling the universities to incorporate enhancements developed inhouse and having the prospect of long time sustainability and prosperity due to the active community supporting the project. In terms of hardware, the IBM GPFS Storage Server (GSS) storage platform was selected. Coordination efforts to enable all participating

universities to join the Shibboleth-based Authentication and Authorization Infrastructure, operated by the German Research Network (DFN-AAI) and required for the envisioned self enrollment portal, were also started at WWU Münster.

Last but not least, a student marketing team was recruited by the WWU Marketing Center to set up an online viral marketing campaign in preparation for the planned public launch, aiming to reach a wide audience mainly through approx. 400 Facebook groups at the participating universities. For better brand recognition, the easy to memorize neologism "sciebo" (short for "science box") was coined by the ZIV's public relations team, along with trademark and domain registration and the conception of the sciebo elephant logo (Figure 2).

Despite an abundance of unexpected last minute challenges, the pre-announced date for public launch on 2 February 2015 could be met.

1.2. Legal Framework of the Consortium

The considerations on the legal framework resulted in the decision that the formation of a dedicated legal entity for the consortium named "Sync & Share NRW" was futile, and that WWU Münster as consortium lead should conduct all legal businesses.

The consortium was established by consortium agreements between WWU Münster and all consortium partners. Additionally, agreements for data processing outsourcing for the operation of the sync & share cloud service were made between WWU Münster and its partners. Common terms of use were provided to the participating universities who are legally the institutions offering the sync & share cloud service to their respective end users.
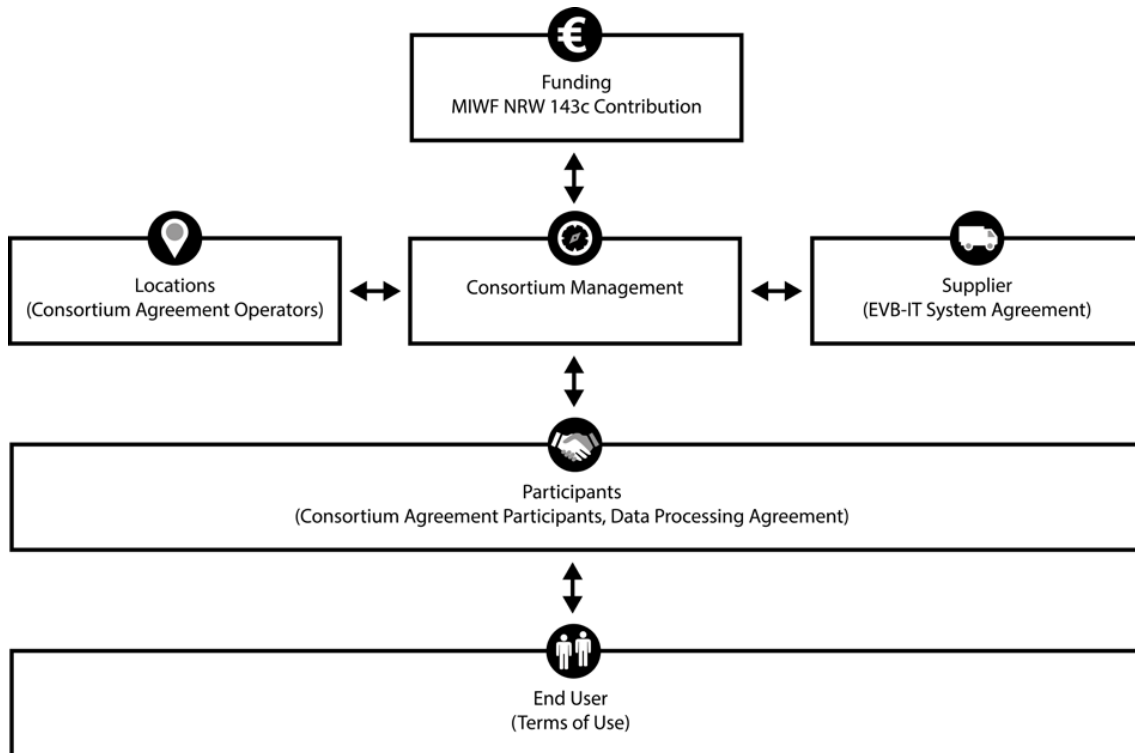
Figure 1: The framework of the Sync & Share NRW consortium.

2. Solution Architecture

2.1. Service Specifications and Features

The Sync & Share NRW consortium has agreed on the following specifications for the sciebo cloud service: First, cloud storage functionalities will have top priority, i.e. at the starting point there will not be any other features apart from user-friendly file up- and downloading, file sharing (e.g. via anonymous links), and continuos synchronisation of directories on personal computers. Sciebo personal boxes will have a 30 GB quota; employees can increase their quota to 500 GB if needed by means of a self service tool.

Second, access will be possible via desktop sync clients and apps for all common operating systems, via web interface, and via WebDAV which means setting up sciebo as drive share. Sciebo's user authentication and authorization will be done via Shibboleth/ SAML through the DFN-AAI service, assuring that sciebo accounts will only be created for persons correctly authorized by their home university. There will also be a compulsory re-authorization after six months to clear users no longer eligible for the service. When no timely re-authorization takes place, a six month grace period starts after which access to sciebo will be blocked. All user data will be deprovisioned after another three months.

Figure 2: The sciebo homepage, prominently featuring the sciebo logo.

Third, to manage the support with very limited human resources, a self service web portal will be provided for sciebo users. Also, trouble tickets will be assigned to the helpdesks of the participating institutions to exploit existing support structures, with second-level support by the central sciebo support team and third-level support by software supplier ownCloud for the Enterprise version license (purchased including five years of support).

Fourthly, in matters of availability there will be no explicit SLAs between hosting institutions and the participating universities. Instead, the consortium has agreed on a target availability of 99.5% p.a. for the sciebo service, with a monthly availability above 98.0% and a maximum continuous service outage of four hours, deemed feasible and realistic based on data center availability records of WWU Münster. Server and storage hardware will be a highly redundant setup with no single point of failure. Additionally, the GSS storage system will operate with triple-parity RAID 8+3 and is thus extremely unlikely to loose data. Also, snapshots will be taken daily and stored for 14 days as precaution against operational errors that could lead to data loss.

2.2. System architecture and integration

From the beginning of the Sync & Share NRW project it was clear that the hardware platform was to be distributed over several university data centers in NRW due to the following reasons:

• rack space and power supply need,

• internet bandwidth that could be dedicated to the operation of sciebo, and

• as a symbol for the multi-university cooperative effort.

The chosen software solution had to be able to realize this scenario in one cloud service with sharing of data being possible between all users, regardless of the actual storage location of their data - this feature being one of the decisive factors for ownCloud. The Universities of Bonn, Duisburg-Essen and Münster were selected as locations for the sciebo sites, each hosting individual ownCloud instances for every single participating university (Figure 3).
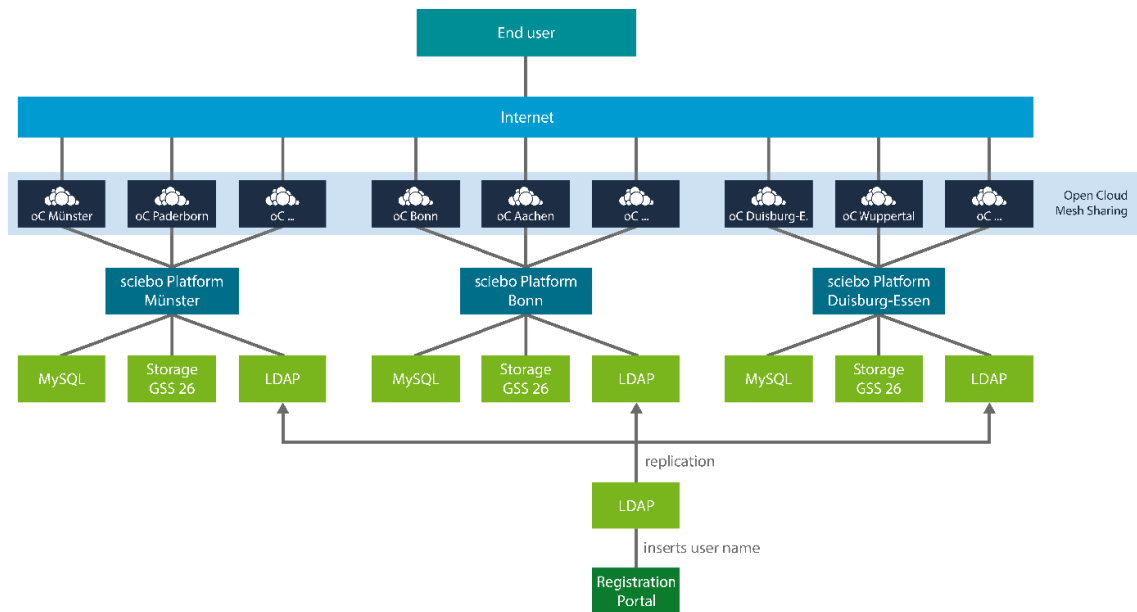

Figure 3: Schematic view of the multi site sciebo architecture.
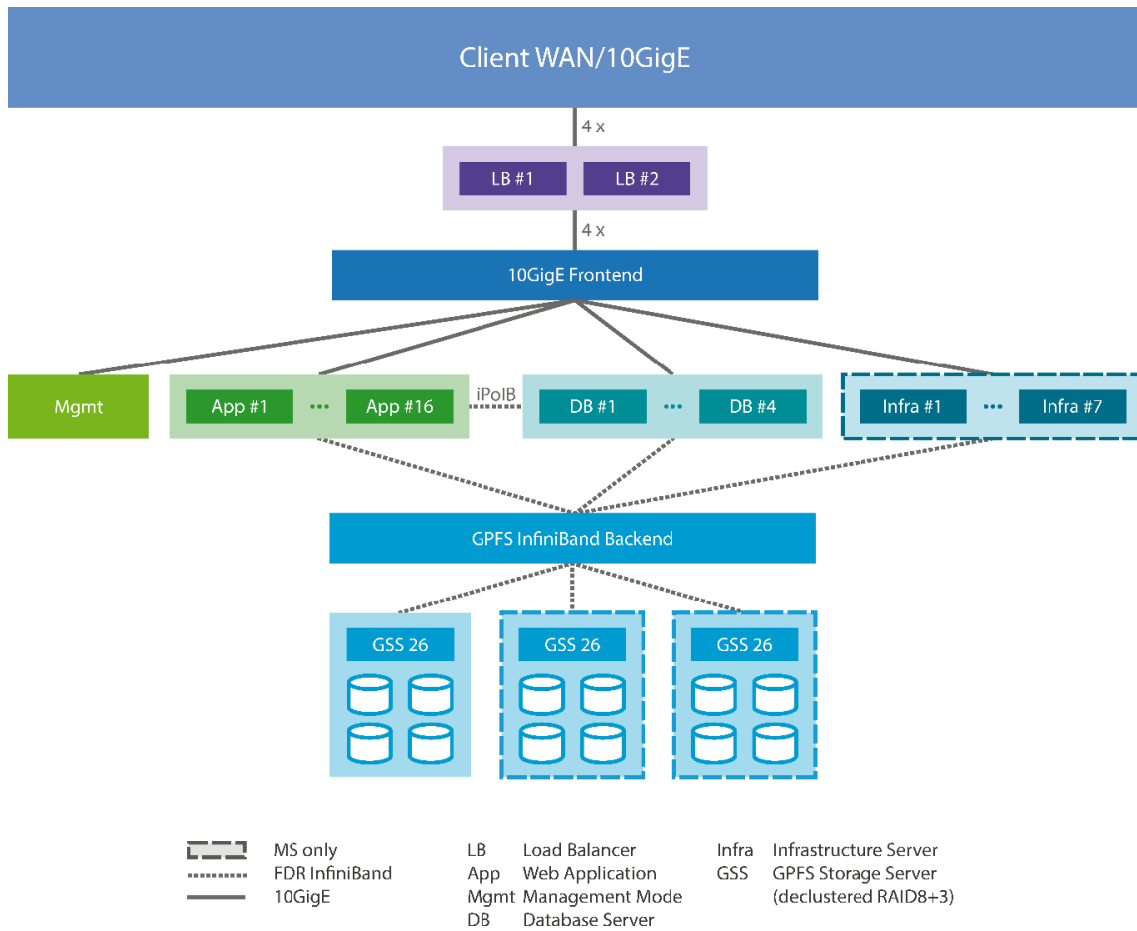
Figure 4: System architecture for a sciebo site installation.

The hardware platform at the three sites is virtually identical (Figure 4).The software stack is comprised of a wide range of production quality open source tools (Figure 5).
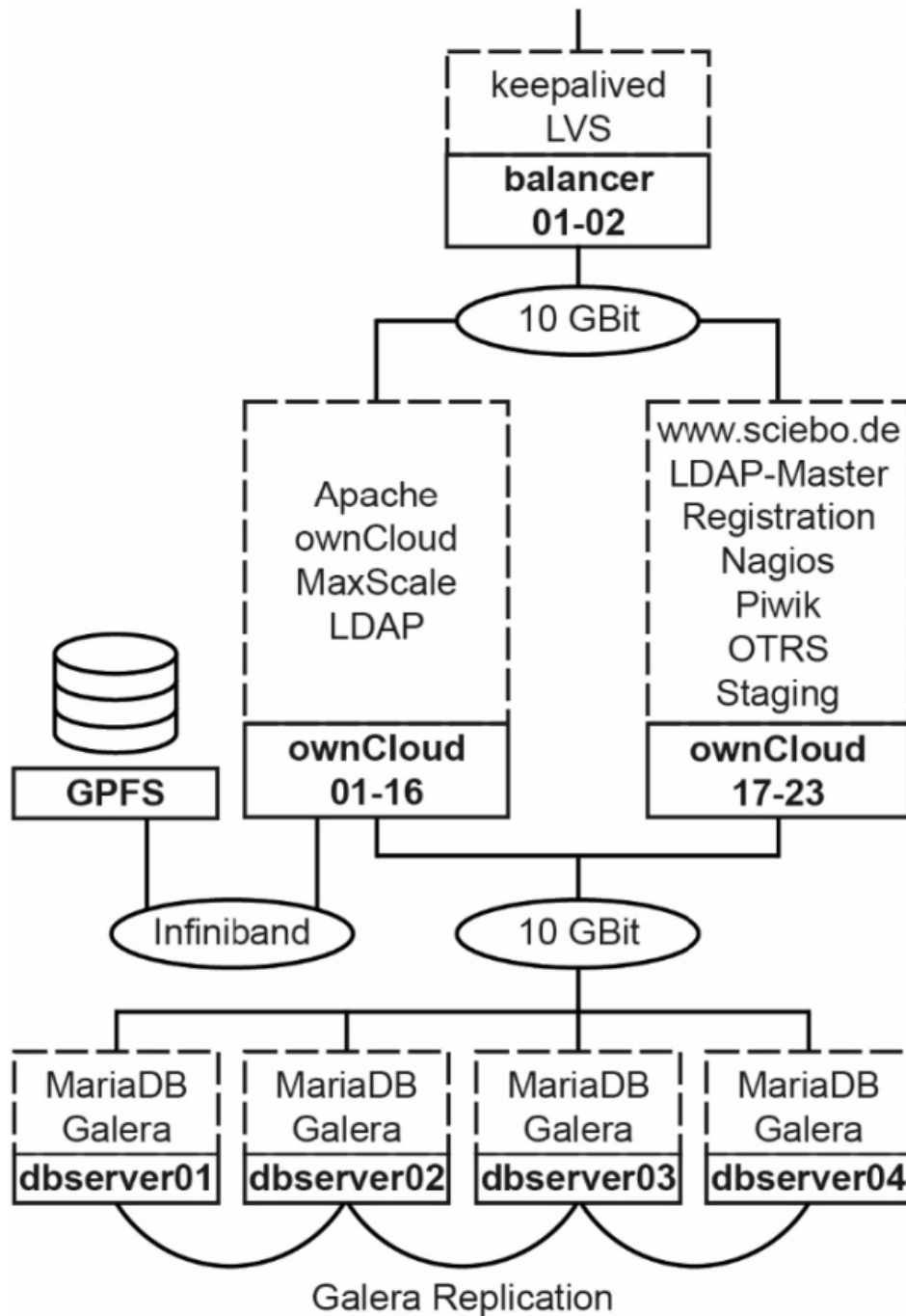
Figure 5: Details on the software tools used in the sciebo installations.

All user accounts are created via the self enrollment portal (with authentication and authorization via DFN-AAI) and maintained in an LDAP database replicated between the three sites. Sharing data between sciebo users from different universities (and thus between different "ownClouds") is done by means of the server-to-server sharing mechanism (also known as Open Cloud Mesh) which allows for the creation of this "cloud of clouds" and was developed by ownCloud in view of the Sync & Share NRW project.

3. First Results and Next Steps

Of the 22 universities committed to the project, 15 managed to overcome all technical (DFN-AAI integration of the respective identity management systems) and organisational (signing of contracts in university presidia, consent of personell representatives, etc.) challenges on time for the launch, with the others subsequently joining in. Further universities and research institutions not yet committed are now considering to join the consortium.

After just half a year of operation, 25,000 users have signed up for sciebo (Figure 6), the total data volume stored has already transgressed 60 TB and is rapidly increasing - in good agreement with the Theory on Diffusion of Innovation (Rogers, 2003) and proving the effectiveness of the online marketing campaign. To gain further user acceptance, it is essential to continually develop the service based on the users' wishes. Therefore, more studies are planned.

The establishment of advisory boards is supposed to further increase users' confidence: In addition to a scientific advisory board, which particularly focuses and stimulates the accompanying research, an IT-security and legal affairs advisory board has been established, which is comprised of representatives from the following groups: users (staff councils, student representatives, university administrations), data protection officials, IT security experts and scientists in the field of IT law. A certification of the service according to a recognized standard (e.g. ISO 27001, BSI IT Baseline Protection) is also currently under discussion.
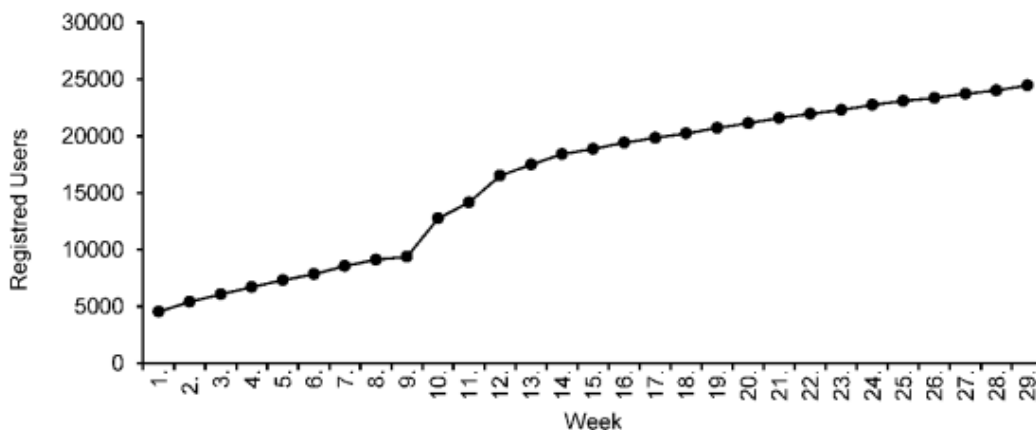

Figure 6: Diffusion of sciebo.

Since we perceive sciebo not just as a sync & share cloud storage service but as the seed for a versatile information infrastructure (Pipek et al, 2009), further research will especially focus on user demand for and adoption of novel usage scenarios based on a powerful, elastic, well established and widely used cloud platform. Among these additional usage scenarios, research data management and e-learning are the most concrete. For research data management, we see huge potential for addressing the collaboration domain in the curation domain model (Treloar et al. 2007, Klump 2011) and creating further benefits for researchers by providing easy to use interfaces to the publication domain. In the field of e-

learning, close integration with the most common e-learning systems at NRW universities (Moodle, Ilias) are planned, making it possible to easily distribute digital materials to student mobile devices for paperless learning.

4. Conclusion

The introduction of a service for up to 500,000 users at more than 20 highly heterogeneous universities, with a minimum of human resources, a relatively unknown technology and with strong competition from commercial providers, was often viewed with skepticism and considered an impossible task. Today - less than three years after the initial idea - these voices have been silenced. Due to the extensive preparation, the use of external expertise, the lively exchange with other similar projects and the full inclusion of user feedback, the project has been very successful so far. The strong growth of the user base, the positive coverage in reputable media and, especially, numerous additional requests for participation serve as proof of this.

5. References

Davis, F.D. : Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. In: MIS quarterly, 13 (3), 1989, pp. 319-340.

Dropbox, TechCrunch & The Next Web: Anzahl der Dropbox-Nutzer weltweit zwischen Januar 2010 und Mai 2014 (in Millionen). 2014. http://de.statista.com/statistik/daten/studie/326447/umfrage/anzahl-der-weltweiten-dropbox-nutzer/ (last check 2015-10-30)

Eurostat: Anteil der Nutzer von Cloud-Diensten nach Ländern in Europa im Jahr 2014: http://de.statista.com/statistik/daten/studie/381271/umfrage/nutzung-von-cloud-diensten-durch-einzelpersonen-in-europa-im-laendervergleich/ (last check 2015-10-30)

Klump, J.: Langzeiterhaltung digitaler Forschungsdaten. In: Büttner, S., Hobohm, H.-C., Müller, L. (Eds.): Handbuch Forschungsdatenmanagement. Bock + Herrchen, Bad Honnef, 2011, pp. 115–122.

Meske, C.; Stieglitz, S.; Vogl, R.; Rudolph, D.; Öksüz A.: Cloud Storage Services in Higher Education - Results of a Preliminary Study in the Context of the Sync&Share-Project in Germany. In: Zaphiris, P. Ioannou, A. (Ed.): Learning and Collaboration Technologies. Designing and Developing Novel Learning Experiences. Lecture Notes in Computer Science. (Proceedings of the 16th International Conference on Human Computer Interaction (HCI International) 2014. Crete, Greece. Cham: Springer International Publishing, pp. 161-171.

Pipek, V.; Wulf, V.: Infrastructuring: Toward an Integrated Perspective on the Design and Use of Information Technology. In: Journal of the Association for Information Systems, 10 (5), 2009, Article 1.

Rogers, E. M.: Diffusion of Innovations (5th ed.). Free Press, New York, 2003.

Stieglitz, S.; Meske, C.; Vogl, R.; Rudolph D.: Do Universities Need To Host a Cloud Computing? Proceedings of the International Conference on Information Systems (ICIS), 2014, Auckland, New Zealand.

Treloar, A.; Groenewegen, D.; Harboe-Ree, C.: The Data Curation Continuum: Managing Data Objects in Institutional Repositories. In: D-Lib Magazine, 13 (9/10), 2007, pp. 13.

Vogl, R.; Angenent, H.; Bockholt, R.; Rudolph, D.; Stieglitz, S.; Meske, C.: Designing a Large Scale Cooperative Sync&Share Cloud Storage Platform for the Academic Community in North Rhine-Westfalia. In: Sukovski U. (ed.): ICT Role for Next Generation Universities - 19th European University Information Systems - EUNIS 2013. Congress Proceedings, Riga: Riga Technical University, 2013, pp. 205-208.