GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

ArCHO

Eine Virtuelle Forschungsumgebung im Spannungsfeld von Open Access, Nachhaltigkeit und Datenschutz

Dhd Bern, 17.2.17

Felix Lange, Urs Schoepflin, Dirk Wintergrün (MPIWG), Oliver Wannewetsch (GWDG)

Einleitung

Projektkontext

- ArCHO: Archiv Cultural Heritage Online
    - Übergreifendes Ziel : Wissenschaftliche Voraussetzungen für eine Infrastruktur zur Langzeitspeicherung von Forschungsdaten
    - Kooperation des Forschungsprogramms zur Geschichte der Max-Planck-Gesellschaft am MPI für Wissenschaftsgeschichte mit der
      Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG)

Einleitung

Projektkontext

- Forschungsprogramm GMPG
    - Geschichte der Max-Planck-Gesellschaft in ihren zeit- und wissenschaftshistorischen Zusammenhängen
    - Geschichte der MPG als wichtiger Teil der kulturellen, politischen und ökonomischen Geschichte der Bundesrepublik im Zusammenhang europäischer und globaler Entwicklungen
- gmpg.mpiwg-berlin.mpg.de

Einleitung

Quellen - Methoden

- Quellengrundlage
  - Veröffentlichungen der MPG
  - Interviews mit Zeitzeugen
  - Verwaltungsschrifttum, Korrespondenzen, Vor- und Nachlässe im Archiv : **> 3 Regalkilometer Akten**

- Anwendung quantitativer Forschungsmethoden. Beispiele:
  - Volltext- und semantische Suchen über größere Quellenbestände
  - Kookurrenzen
  - Netzwerkanalysen
  - Geo-Mapping
  - Topic Modeling

Einleitung

Quellen -Digitalisierung

- Digitalisierungsvorhaben im Projekt: Ca. 20.000 Archivalien
- Zitierfähigkeit muss bis zehn Jahre nach dem Projekt erhalten bleiben.
- Spezifische datenschutzkonforme Aktenzugangsregeln für das Projekt

Einleitung

GMPG - Quellendigitalisierung

- Quellengrundlage
  - Veröffentlichungen der MPG
  - Interviews mit Zeitzeugen
  - Verwaltungsschrifttum, Korrespondenzen, Nachlässe im Archiv :
    > 3 Regalkilometer Akten

- Zwei große Digitalisierungsvorhaben im Projekt
  - On-Demand durch studentische Hilfskräfte
  - Massendigitalisierung ab 2017: ca. 20.000 Akten

Datenschutz

Sonderfall Zeitgeschichte

- Interessenabwägung Forschung - Persönlichkeitsrechte der Betroffenen
  - Informationelles Selbstbestimmungsrecht (GG Art. 2 u.1) - Forschungsfreiheit (GG Art. 5)

- Generelles Datensparkeitsgebot, „Übermaßverbot“ (BDG ...)

- Zweckbindung der Nutzung – Erlaubnisse erlöschen mit Projektende

- Beachtung regionaler, Sonderregeln: Je nach Archivstandort gelten ggf. Landesdatenschutzgesetze

Datensparsamkeitsgebot versus Big Data

Datenschutz

Open Access

- MPIWG ist der „Berlin Declaration on Open Access“ verpflichtet
- Daten sollten – z.B. für komparative Studien – der Forschung erhalten bleiben
- Undenkbar, 20.000 Aktendigitalisate nach Projektende zu vernichten oder praktisch unbenutzbar zu machen

Datensparsamkeitsgebot versus Open Access

Datenschutz

Regelungen in GMPG

- Eigene Aktenzugangsregeln
- Sicherheitsstufenmodel
  - I Öffentliches Material (aber: Urheberschutz!)
  - II Archivmaterial (für das Projekt zugänglich)
  - III Datenschutzsensibles Material
    - Einzelbeantragung, Vorsichtung durch Projektkollegium, detailiertes Zugangsprotokoll
  - Abgestufte Zugangsrechte: Kollegium, Forscher, IT-Personal (intern), IT-Personal (extern), Hilfskraft

Infrastrukturen – Archivierung - Datenschutz

Verfahren in wissenschaftlichen Archiven

- Archive sind traditionelle Gatekeeper zu wissenschaftlichen Quellen und garantieren die Einhaltung von Datenschutzstandards
- In der Diskussion: Virtueller Lesesaal versus Digitaler Lesesaal
- Dauerhafte Archivierung von Projektdaten (GMPG: ~500 TB) ?
- Bereitstellung von rechtssicheren Virtuellen Forschungsumgebungen ?

Infrastrukturen – Archivierung - Datenschutz

Beispiel Sozialwissenschaften

- Wunsch nach „Sekundärauswertung“ nach Ende von Studien
- Ständiger Ausschuss Forschungsdateninfrastruktur
- Große verstetigte Einrichtungen als Daten-Hubs
  - Secure Data Center des GESIS Leibniz-Institut
    - „Safe Room“
    - Ausgewählte lokale Analysesoftware in VM

Infrastrukturen – Archivierung - Datenschutz

Situation in den Geisteswissenschaften

- Viele Initiativen und Regelwerke: Nestor, HDC, DARIAH, WissGrid, ...
- Vertraulichkeit / Nachhaltigkeit (nestor kap. 5.3./Steinmetz 2002) :
  Integrität | Vertraulichkeit | Verfügbarkeit
- Vorwiegendes Verfahrensmodell : „Abgabe“ der Daten nach Projektende, Anwendungskonservierung (z.B. HDC)
- Beobachtung: Vorwiegen „sozialer Regelungen“ des Zugangs während der Projekte

Meist klare Trennung Projektlaufzeit - Archivierungszeitraum

Zwischenfazit

Probleme und Desiderata aus Sicht von GMPG

- Spannungsfeld: *Open Access – Datenschutz – Big Data*
  - *Skalierbarkeit*
  - *Portierbarkeit*

- Bei datenschutzsensiblem Material: Weniger klare Unterscheidungen zwischen Projektlaufzeit und Archivierungszeitraum

- Wunsch, Rohdaten und Derivate hinsichtlich der Zugangsregelungen klar zu unterscheiden => rechtliche Möglichkeiten voll ausschöpfen

Desideratum für GMPG: Eine rechtssichere Langzeitinfrastruktur während *und* nach dem Projekt

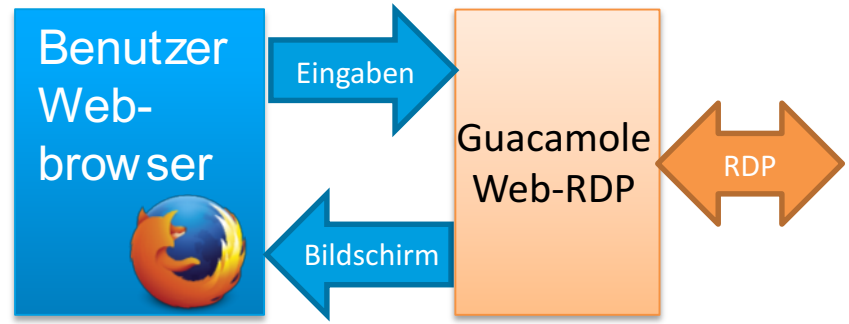
ArCHO

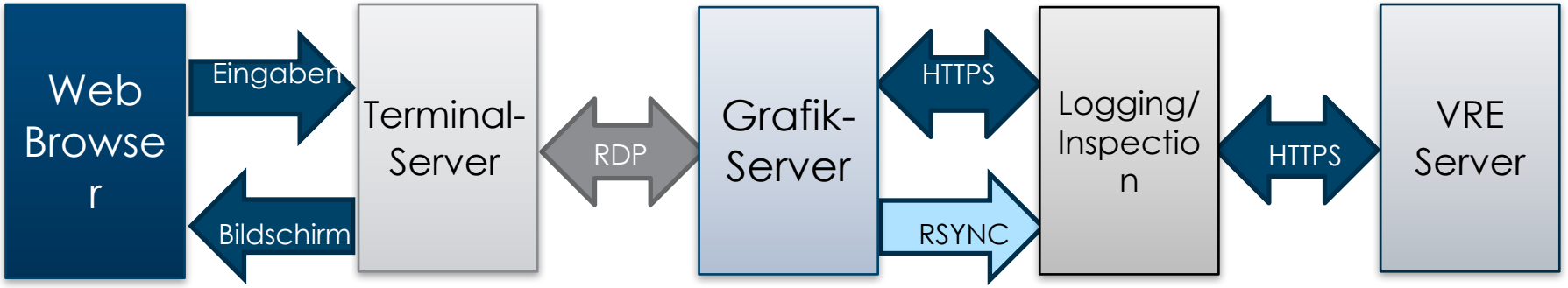
Architekturprinzipien

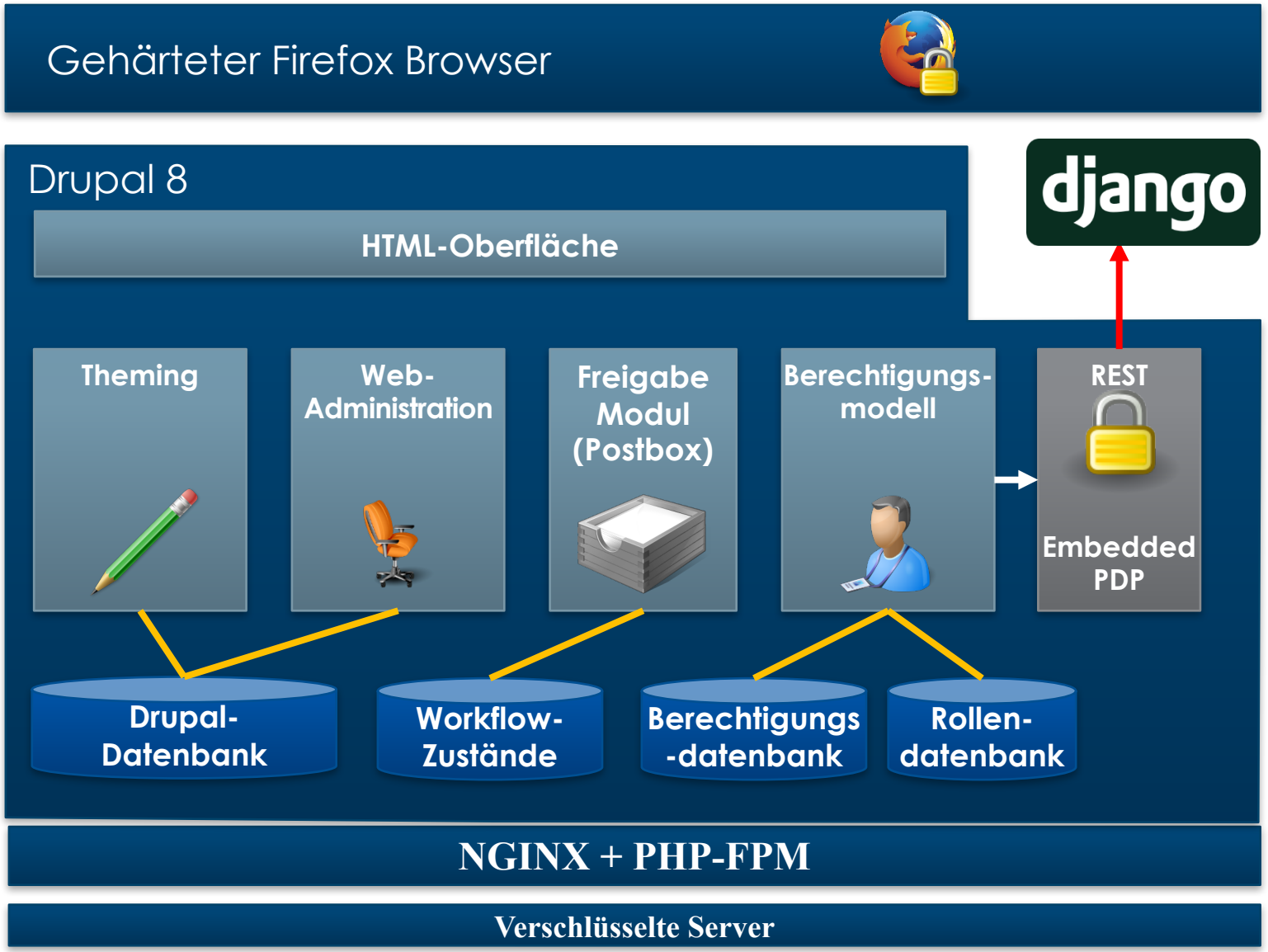
- Konvergenz von Archiv- und Projektinfrastruktur
- Feingranulare Rechteverwaltung
- Sicherheit UND Nachhaltigkeit durch Virtualisierung

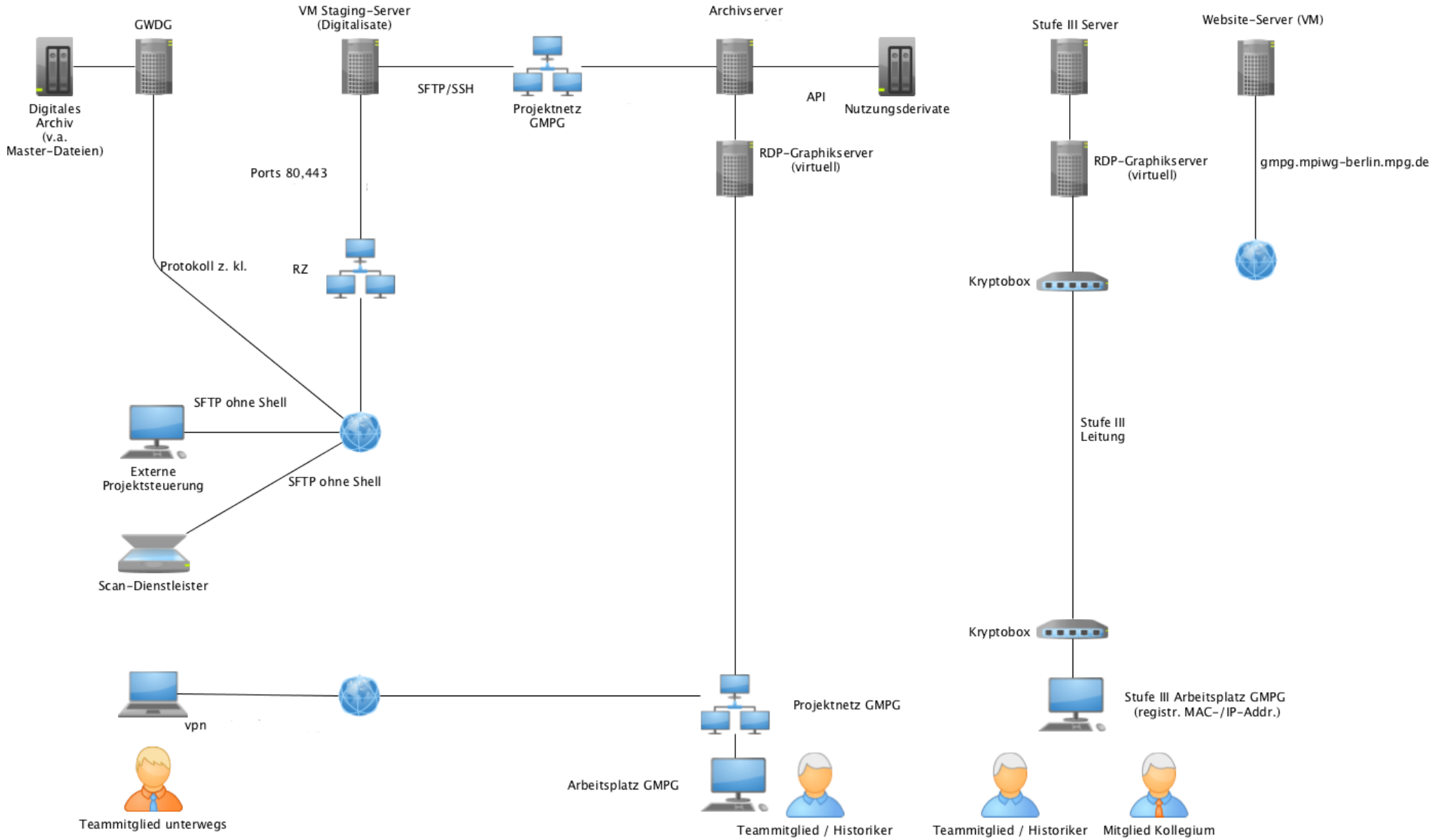
Architektur

Auslieferung von Inhalten mit Guacamole RDP






Gehärteter Firefox Browser
Drupal 8
HTML-Oberfläche
django
Theming
Web-Administration
Freigabe Modul (Postbox)
Berechtigungs-modell
REST
Embedded PDP
Drupal-Datenbank
Workflow-Zustände
Berechtigungs-datenbank
Rollen-datenbank
NGINX + PHP-FPM
Verschlüsselte Server


GWDG
VM Staging-Server (Digitalisate)
Archivserver
Stufe III Server
Website-Server (VM)
Digitales Archiv (v.a. Master-Dateien)
SFTP/SSH
Projektnetz GMPG
API
Nutzungsderivate
RDP-Graphikserver (virtuell)
RDP-Graphikserver (virtuell)
gmpg.mpiwg-berlin.mpg.de
Ports 80,443
Protokoll z. kl.
RZ
Kryptobox
SFTP ohne Shell
Externe Projektsteuerung
SFTP ohne Shell
Stufe III Leitung
Scan-Dienstleister
Kryptobox
vpn
Projektnetz GMPG
Stufe III Arbeitsplatz GMPG (registr. MAC-/IP-Addr.)
Arbeitsplatz GMPG
Teammitglied unterwegs
Teammitglied / Historiker
Teammitglied / Historiker
Mitglied Kollegium

Fazit

Architekturprinzipien für eine rechtssichere, nachhaltige VRE

- Vorteile einer technischen gegenüber einer „sozialen“ Rechteverwaltung
  - Portierbarkeit (z.B. nach Projektende)
  - Skalierbarkeit

- Emulation/Virtualisierung/Container als wichtiges Prinzip in nachhaltigen Infrastrukturen

- Eine vollständige Eigenlösung eines Forschungsprojekt ist unter den geschilderten Bedingungen nicht sinnvoll oder gar unmöglich

Fazit

Verwaltung von Archivdaten durch Forschungsprogramme

- Realisierung eines erweiterten „Virtuellen Lesesaals“ durch Projektgruppen möglich

- Die Frage der künftigen Arbeitsaufteilung
  Forscher – Archive – Rechen-/Datenzentrum ist davon unberührt