Persistent Identifiers für die Wissenschaft vom European Persistent Identifier Consortium (EPIC)

(PID: http://hdl.handle.net/11858/00-ZZZZ-0000-0001-6D1D-0)

Tibor Kálmán
tibor [dot] kalman [at] gwdg [dot] de

Workshop “Archivierung sozial- und wirtschafts-wissenschaftlicher Datenbestände”
(15.09.2011)

Agenda

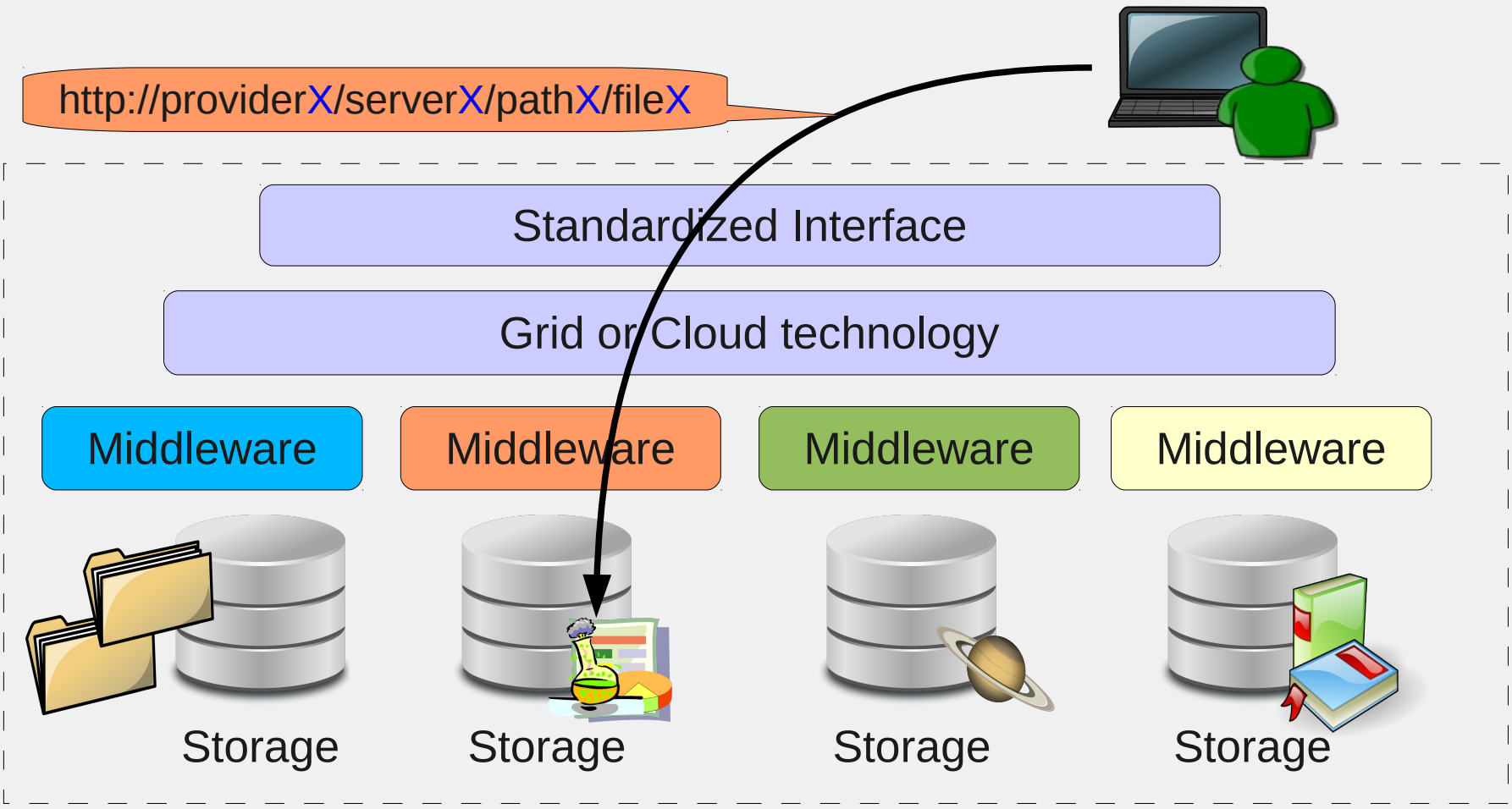
- Persistent Identifiers (PIDs) für die Forschung
- Konsortium für PIDs
- EPIC PID-Service
  - RESTful Web Service
- Vorteile und Voraussetzungen von PIDs
- Ausblick
- Zusammenfassung

Motivation (1)

- Der Umfang der digitalen Objekte wächst stetig in allen Bereichen der Wissenschaften
- Der Zugang und die Langzeitarchivierung sind wichtige Aufgaben
- Zugriff über einheitliche Schnittstelle zum Speicher- und Repositorysystem
  - Das Interface ist Technologieunabhängig gestaltet
  - und somit von der konkret eingesetzten Infrastruktur abstrahiert
    - Mit Hilfe von HTTP, WebDAV
    - Das ermöglicht, dass auf die gespeicherten digitalen Objekte auch mit gängigen Tools zugegriffen werden kann (Web- und Dateisystem-browsern)
  - Ein Austausch zu grundelegenden Speichertechnologie ist ”jederzeit” möglich

Einheitliche Schnittstellen zu digitalen Objekten

- Zugriff zu Forschungsdaten: über einheitliche Schnittstelle



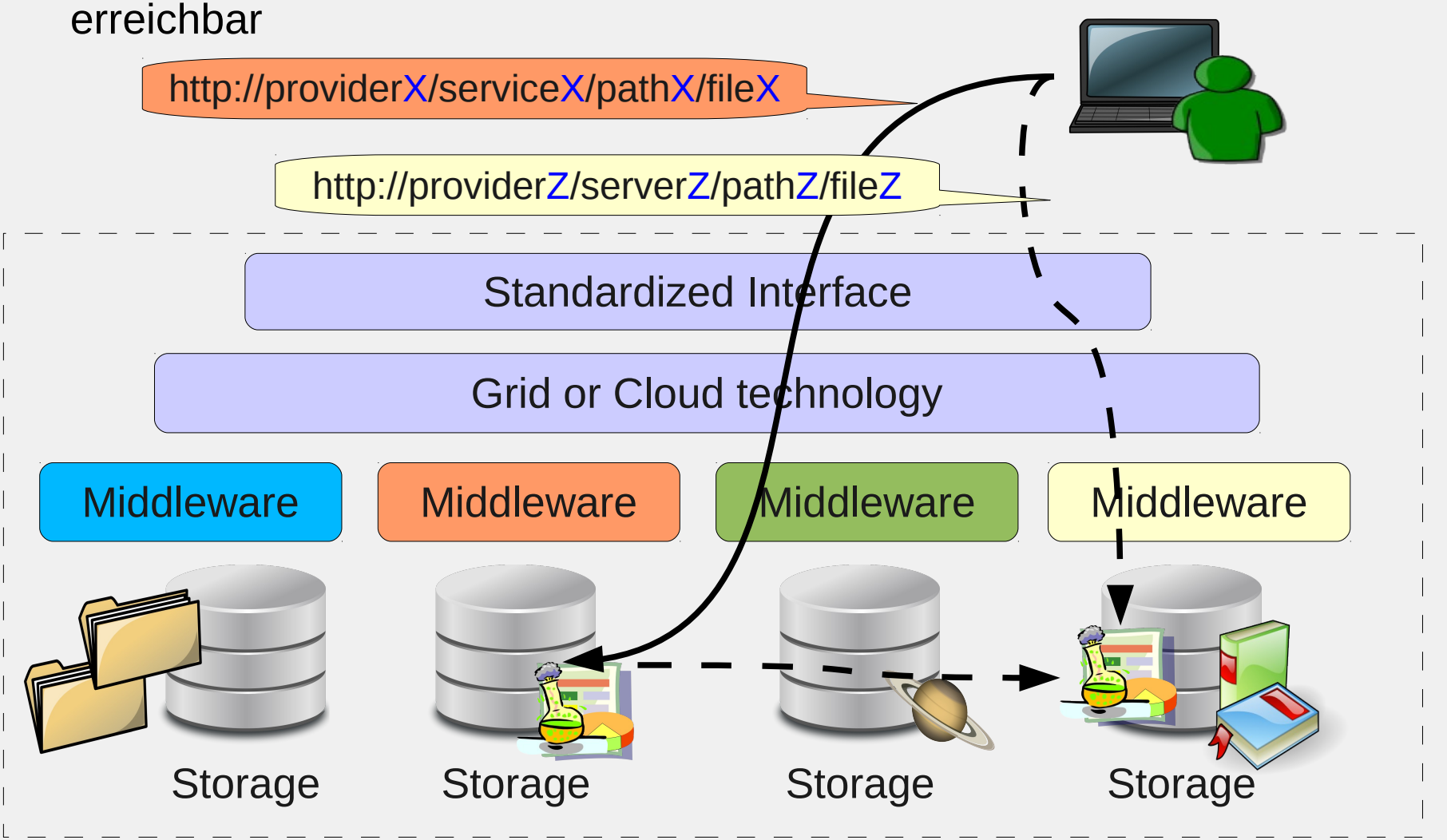
Motivation (2)

- Die Menge der gespeicheren Objekte wächst, wie sollen sie identifiziert werden?
  - Oft wird ein URI dabei benutzt
- Uniform Resource Identifier (URI): die Adresse unter der ein digitales Objekt abgelegt wird
  - Die Adresse ist oft nicht persistent (wird oft nicht dauerhaft erhalten- etwa wegen Datenmigration, usw.)
  - nach jeder Änderung ist das digitale Objekt unter einer neuen Adresse erreichbar
  - Die URI enthält physikalische Pfade und semantische Inhalte

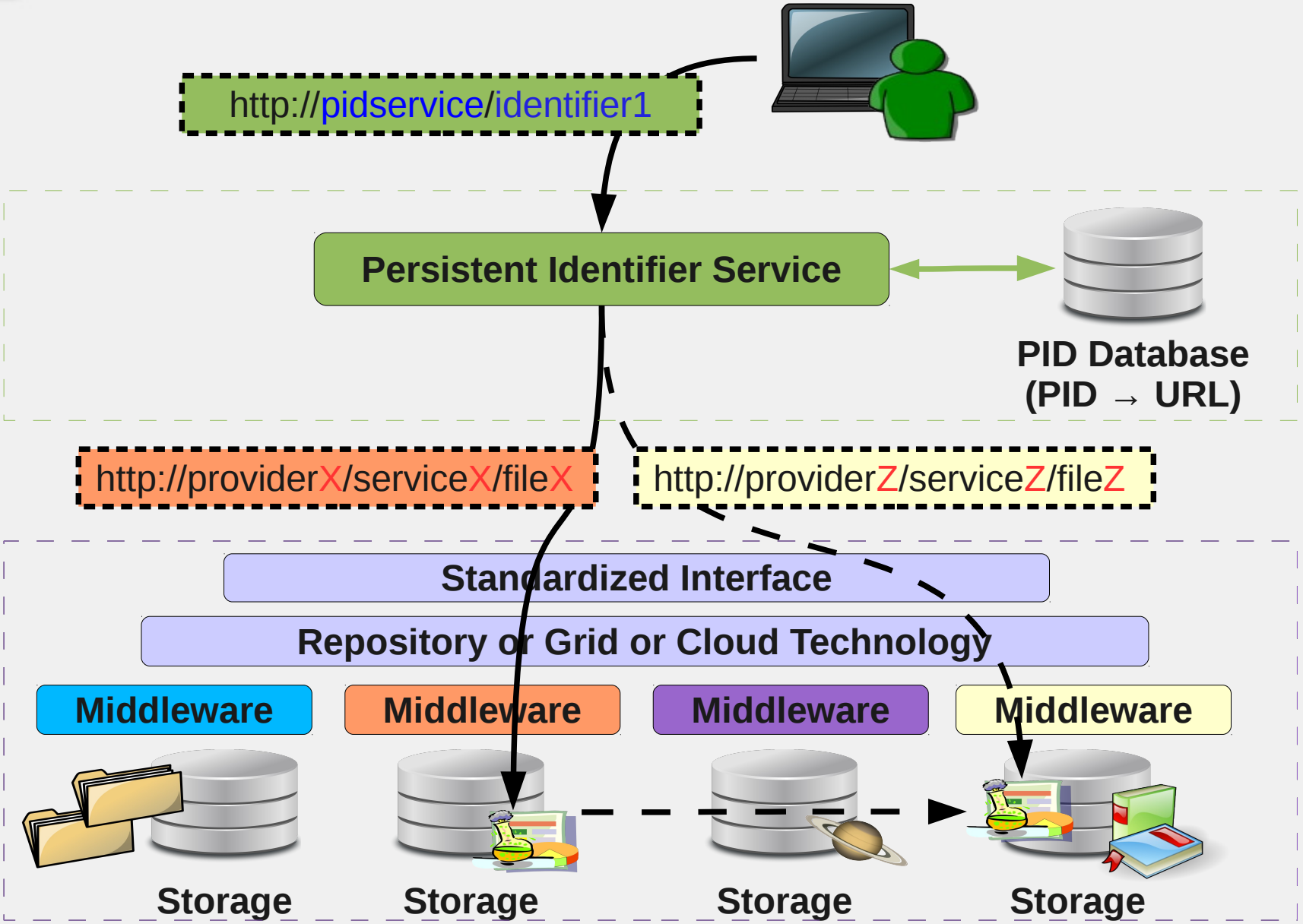
Es ist nicht möglich, die digitalen Objekte langlebig mit URIs zu referenzieren!

Referenzierung mit der physikalischen Adresse

- nach jeder Änderung ist das digitale Objekt unter einer neuen Adresse erreichbar



Referenzierung durch Persistent Identifier

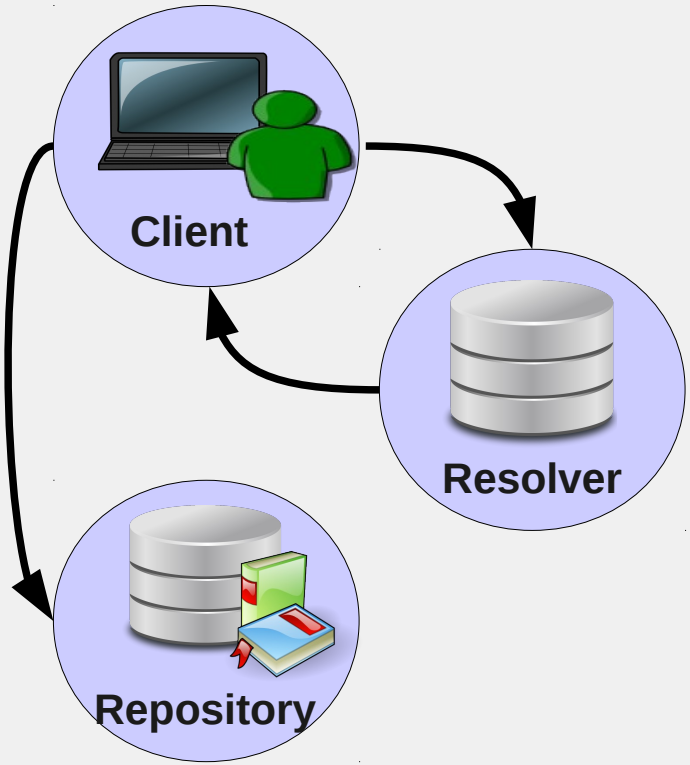


Persistent Identifier (PID)

- Wissenschaftliche Institute brauchen eine Strategie für die Langzeitarchivierung und den dauerhaften und beständigen Zugriff auf Forschungsdaten.
- **P**ersistent **ID**entifier (PID)
  - Digitale Objekte sind im Repositories registriert
  - Mit einem unveränderlichen Identifikator (PID)
  - Die unterliegenden Systeme können sich ändern ("living organisms")
  - Eine Migration ist auf verschiedenen Ebenen möglich (Änderung von Hardware, Software, Format, usw.)
- Für die Vergabe, Verwaltung und Auflösung von PIDs:
  - Ein allgemein vereinbarter Prozess ist nötig
  - Handle System, ähnlich Domain Name System (DNS)

Es ist möglich, die digitalen Objekte langlebig mit PIDs zu zitieren!

Langfristige Referenzierung



Eindeutige und dauerhafte Erreichbarkeit von digitalen Objekten:

- Objekte erhalten PIDs
- PID: Eine neue Schicht in der Infrastruktur
- Resolve:
  - Muss administriert werden
  - Muss mit hoher Verfügbarkeit laufen
  - Muss mit den Archives zusammenarbeiten können
  - Und das alles auch langfristig

Agenda

- Persistent Identifiers (PIDs) für die Forschung
- Konsortium für PIDs
- EPIC PID-Service
  - RESTful Web Service
- Vorteile und Voraussetzungen von PIDs
- Ausblick
- Zusammenfassung

PID Service für die Wissenschaft

- Nicht alle Institute können oder wollen einen eigenen Dienst zum Verwalten von PIDs betreiben
- Seit 2009 läuft bei der GWDG (für die Max Planck Gesellschaft) ein PID-service
- Basierend auf dem Handle System (http://www.handle.net/)
- Ziel: Erzeugen, Verwalten und Auflösen von Identifikatoren von Forschungsdaten (wissentschaftlichen digitalen Objekten)
- Zusammen mit anderen europäischen Partnern wurde ein Konsortium gebildet, um diese Dienstleistungen europäischen Wissenschaftlern zur Verfügung zu stellen
  - European Persistent Identifier Consortium (EPIC)
  - http://www.pidconsortium.eu/

Das EPIC Konsortium

- European Persistent Identifier Consortium (EPIC)
- Mit dem Ziel einen **P**ersistent **ID**entifier (PID) Service anzubieten
- Der Fokus ist die europäische Forschungslandschaft und kulturelle Institutionen
- EPIC besteht aus drei europäischen Rechenzentren:
  - Wird durch nationale Programme finanziert
  - Mit langer Erfahrung mit dem Betrieb von stabilen und hochverfügbaren Diensten
  - Mit der Möglichkeit SLAs anzubieten
  - Sind in diversen eScience Projekten beteiligt

Partner in EPIC: GWDG (1)

- Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
- eine gemeinsame Einrichtung der Georg-August-Universität Göttingen und der Max-Planck-Gesellschaft
- Sie erfüllt die Funktion eines Rechen- und IT-Kompetenzzentrums für die Max-Planck-Gesellschaft und des Hochschulrechenzentrums für die Universität Göttingen.
- GWDG wurde 1970 als gemeinnützige Gmbh gegründet
- Der Standort liegt in Göttingen
- 25,000 Benutzer
- 1000 wissenschaftliche HPC-Benutzer
- Etwa 80 Mitarbeiter

Partner in EPIC: GWDG (2)

- Aufgabenbereiche der GWDG u.a.:
  - High performance computing
  - High performance networking
  - Infrastructure services
  - IT consulting
- Partner in diversen eScience & Grid Projekten:
  - DARIAH-DE, CLARIN-D, D-Grid DGSI
- Projektkoordination in folgenden Forschungsprojekten:
  - Instant-Grid, OptiNum-Grid
- Weitere Projekte: GoeGrid, Kopal, usw.

Partner in EPIC: SARA

- Stichting Academisch Rekencentrum Amsterdam (SARA)
- SARA unterstützt Forscher in den Niederlanden und arbeitet eng mit akademischen Partnern zusemmen sowie öffentlichen und wirtschaftlichen Einrichtungen
- SARA bietet – seit 40 Jahren – an:
  - High performance computing
  - Visualizierung
  - High performance networking
  - Infrastructure services
- Standort in Amsterdam

Partner in EPIC: CSC

- IT Center for Science Ltd (CSC)
- CSC bietet als Partner der finnischen Forschungsinfrastruktur hochqualifizierte IT Dienstleistungen
- Bietet Finnlands leistungsfähigste Supercomputer
- CSC wurde 1970 gegründet und wird seit 1993 als gemeinnütziges Unternehmen fortgeführt
- Standort in Espoo, nahe zum Otanie Campus der Helsinki University
- Über 180 Mitarbeiter
- 3000 Forscher nutzen die CSC’s Rechenangebote

EPIC Nutzergruppen

- MPG, Max Planck Society

- CLARIN, Common Language Resources and Technology Infrastructure

- DARIAH-DE, Digital Research Infrastructure for the Arts and Humanities

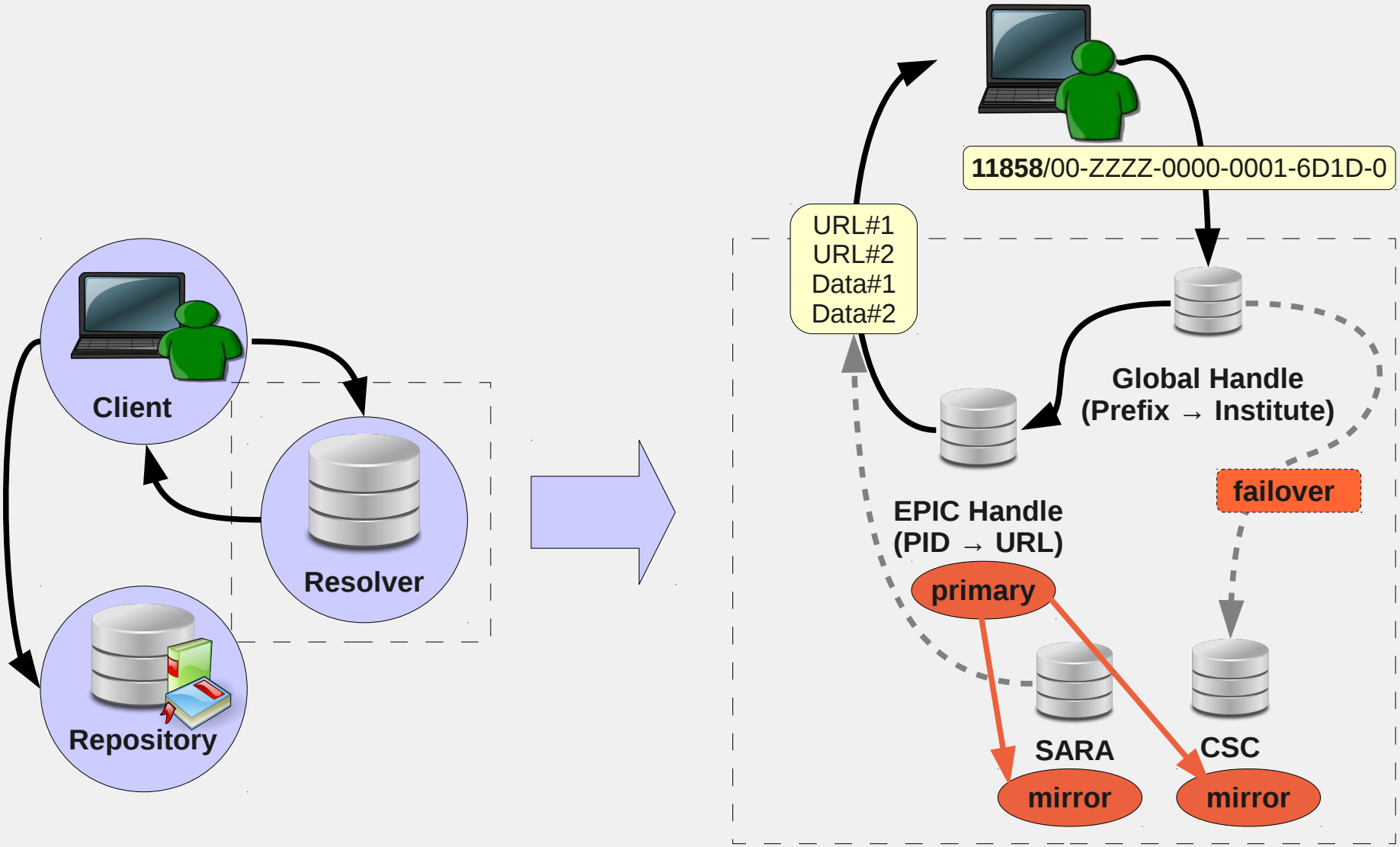
- SUB, Niedersächsische Staats- und Universitätsbibliothek Göttingen

- DKRZ, German Climate Computing Center

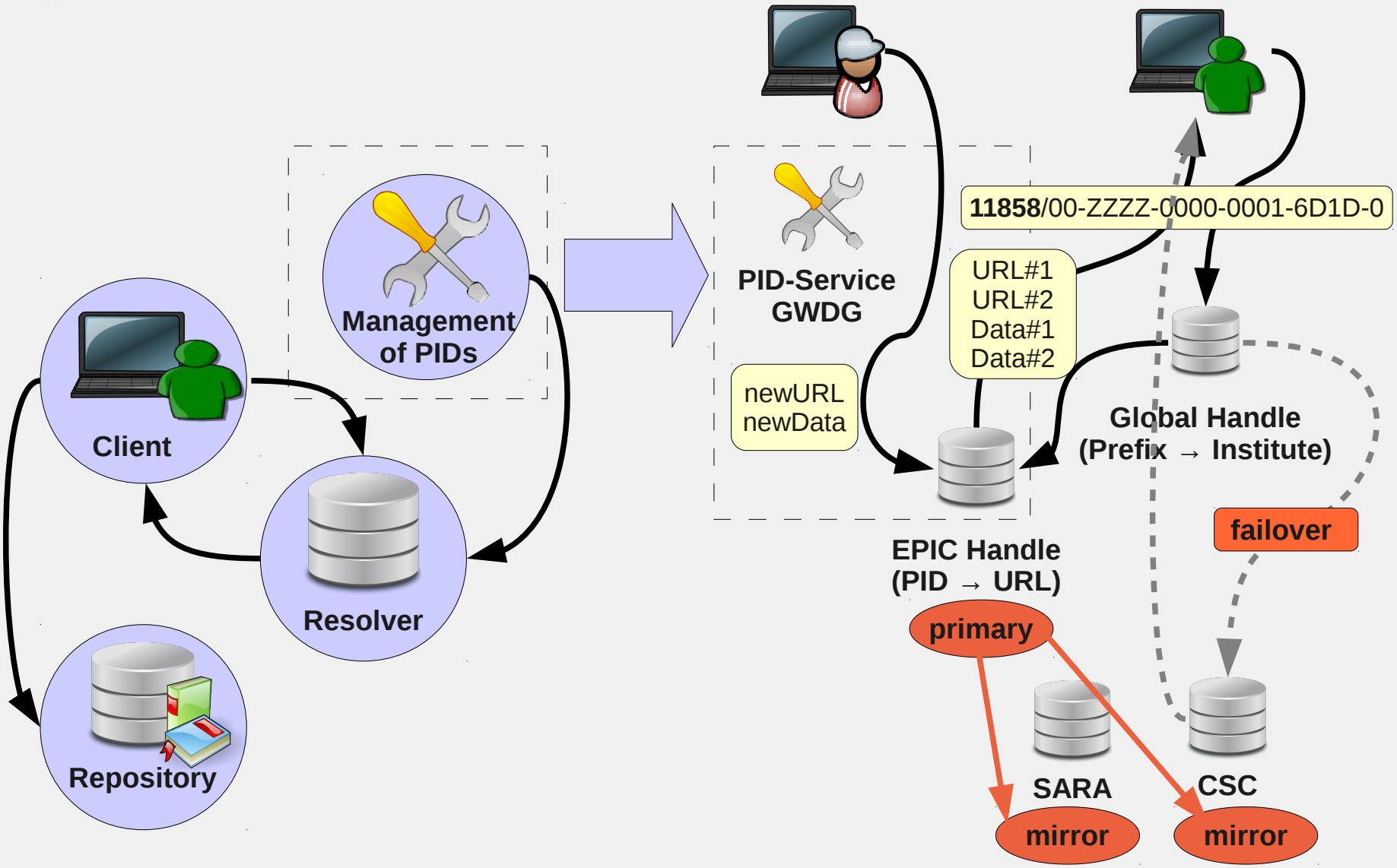
Agenda

- Persistent Identifiers (PIDs) für die Forschung
- Konsortium für PIDs
- EPIC PID-Service
  - RESTful Web Service
- Vorteile und Voraussetzungen von PIDs
- Ausblick
- Zusammenfassung

Auflösung (Resolution) von PIDs in EPIC



Verwaltung von PIDs in EPIC



Zuverlässigkeit und Nachhaltigkeit des EPIC PID-Services

- Basiert auf dem Handle System
  - Das Handle-System wird von vielen Organisationen benutzt
  - Die grundlegende Technologie des Handle-Systems existiert seit ca. 20 Jahren
  - Es ist ein verteiltes System (Global und Lokal Server)
    - Sehr gut skalierbar
- Ein globaler Handle Server für Europa wird bei der GWDG betrieben
  - Verfügbarkeit
  - Sicherheit und Unabhängigkeit
- Die Partner-Organisationen in EPIC haben langfristige und nachhaltige Planung und Finanzierung

Die Syntax von Handles (GWDG) (1)

- Zum Auflösen des PIDs benötigt man einen definierten Prozess
  - Weltweites Handle (PID) Framework
- GWDG benutzt den prefix 11858
  - http://handle.gwdg.de:8000/
  - Dieser ist in das generelle und weltweite Handle Framework integriert..
  - This is integrated into the general and worldwide handle framework
  - Der globale Handle Service deligiert alle Anfragen zur Auflösung mit dem prefix 11858 an die GWDG weiter

Die Syntax von Handle (GWDG) (2)

Der Handle dieser Präsentation:

- http://hdl.handle.net/**11858/00-ZZZZ-0000-0001-6D1D-0**
- **RESOLVER / prfix / fg – inst – num1-num2-num3 - c**
- Die Bedeutung der Felder:

- **prfix** is the handle prefix,
  - which is fixed to 11858
- **fg** is a uppercase hexadecimal flag,
  - can be used for special purposes,
  - to be defined later(derived handles etc)
- **inst** is a field with alphanumerical uppercase digits
  - describes the institution responsible for registration of the handle,
- **num1-num2-num3** 12 digits
  - coded in uppercase hexadecimal digits with delimiters
- **c** is a checksum
  - to ensure plausibility of the handle string.

Schnittstellen zum Erzeugen, Verwalten und Auflösen von PIDs in EPIC

- Aufgabe: Erzeugen, Verwalten und Auflösen von PIDs
- Schnittstellen:
  - (Native) Handle Interface
    - [RFC3652] Handle System Protocol
    - [RFC3651] Handle System Namespace and Service Definition
  - Webinterface:
    - REST-basiertes Web Services Interface
    - http://handle.gwdg.de:8080/pidservice/

Das Webinterface vom PID-Service

- Zugriffsbeschränkungen (Access control):
  - Suche und Auflösung: freier Zugriff
  - Erzeugung neuer PIDs und Verwaltung von existierenden PIDs: Authentifizierung und Authorisierung
- Basic auth: Benutzername + Passwort (genau wie in Web)
- RESTful Web Services:
  - Representational State Transfer (REST), Dissertation von Roy Fielding
  - REST ist eine Softwareachitektur, welche das Internet als Platform für verteiltes Rechnen nutzt
- Ein REST-Beispiel: Ist Dokument-Austausch durch Email schon REST?

REST Design Principles

Grundlegende Designprinzipien von REST:

1) Nutze standard HTTP Methoden
   - Create (POST), Read (GET), Update (~~PUT~~ POST), ~~Delete (DELETE)~~
2) Sei zustandslos (stateless)
   - Abfragen enthalten alle Daten
3) URIs sollten selbsterklärend sein
   - http://handle.gwdg.de:8080/pidservice/write/create
   - http://handle.gwdg.de:8080/pidservice/write/modify
   - http://handle.gwdg.de:8080/pidservice/read/view
4) Klienten wählen das Datenformat
   - XML, JSON, HTML

Agenda

- Persistent Identifiers (PIDs) für die Forschung
- Konsortium für PIDs
- EPIC PID-Service
  - RESTful Web Service
- Vorteile und Voraussetzungen von PIDs
- Ausblick
- Zusammenfassung

Persistenz der Daten vs. Identifikatoren

- Steigendes Datenaufkommen in den Wissenschaften
- Ein Beispiel für einen möglichen wissenschaftlichen Arbeitsablauf:
  - Wissenschafler erzeugen eine große Anzahl von experimentellen Daten
  - Die Wissenschaftler kennen die Bedeutung der Daten erstmal nicht
  - Die langfristige Referenzierbarkeit ist dabei wichtig: ein langfristiger Identifikator wird für die Referenzierung gebraucht
  - Metadaten können extrahiert und mitreferenziert werden
  - Die Daten selbst können repliziert oder zu einem anderen Standort bewegt werden
- Ein PID ist persistent!
  - Seine Gültigkeit kann geprüft und eingegrenzt werden
    - z.B. wenn das Objekt nie referenziert wird (automatisch prüfbar)

Die Referenzen (Identifikatoren) und ihre Metadaten können länger existieren als die wissenschaftlichen Daten selbst

Vorteile von PIDs

Vorteile von PIDs:

- Die Referenzen können länger existieren als die Daten selbst
- Die Referenzen können Daten und ihre Metadaten verbinden
- Einfache Zitierbarkeit für kollaborative Arbeit
- Einfache Referenzierung für Archive oder Replikation

Voraussetzungen von PIDs

Voraussetzungen zur Nutzung von PIDs:

- Die PIDs sollen ein Teil des wissentschaftlichen Arbeitsablauf sein
- Granularität soll eine wissentschaftliche (und nicht technische) Frage sein
  - Kosteneffizienz ist wichtig, denn viele PIDs können ohne großen finanziellen Aufwand genutzt werden
- Der Preis soll von der Wissenschaft bestimmt werden
- Verfügbarkeit
- Sicherheit (insbesondere die Auflösung)

Agenda

- Persistent Identifiers (PIDs) für die Forschung
- Konsortium für PIDs
- EPIC PID-Service
  - RESTful Web Service
- Vorteile und Voraussetzungen von PIDs
- Ausblick
- Zusammenfassung

Ausblick

- Die EPIC API v1 ist sehr einfach (tut was sie tun soll)
- Neu Funktionen sind gewünscht:
  - Neue Anforderungen der Communities, z.B.:
    - 1 PID mit mehreren URLs
    - 1 create request für mehrere PIDs
    - und weitere... :)
  - Auch das Handle-System wurde weiterentwickelt (v7.0)
- Die EPIC API v2 soll bald kommen :)
  - Die EPIC API v2 wird eine generische API sein
  - Wird vom EPIC Konsortium gestaltet, entwickelt und implementiert
- “Politische Unabhängigkeit”: Patent-Besitzer, Domain-Besitzer: ITU
- Erweiterung vom EPIC Konsortium

Agenda

- Persistent Identifiers (PIDs) für die Forschung
- Konsortium für PIDs
- EPIC PID-Service
  - RESTful Web Service
- Vorteile und Voraussetzungen von PIDs
- Ausblick
- Zusammenfassung

Zusammenfassung

- PIDs werden in der Forschung benötigt und sollten Teil der Startegie zur Langzeitarchivierung und den nachhaltigen Zugriff auf Foschungsdaten werden.
- Die PIDs sollen ein Teil des wissentschaftlichen Arbeitsablauf sein
- EPIC ist ein Konsortium, welches einen PID-Service der europäischen wissenschaftlichen Community zur Verfügung stellt
- Der EPIC PID-Service:
  - EPIC API v1 ist eine Schnittstelle mit einem RESTful Web Service
  - Python, Java, Shell Beispiele sind verfügbar
  - EPIC API v2 wird eine mehr generische API sein

Danke!

- EPIC: http://www.pidconsortium.eu/
- PID-Service: http://handle.gwdg.de:8080/pidservice/
- Ein technischerer Vortrag zum EPIC PID-Service: http://hdl.handle.net/11858/00-ZZZZ-0000-0001-6D1D-0