EWIG geht es weiter

Ein Erfahrungsbericht

Ingo Kirchner

Institut für Meteorologie
Freie Universität zu Berlin

4. November 2014

Freie Universität Berlin

Inhaltsverzeichnis

1. Forschungsdatenmanagement
2. Das EWIG Projekt
3. MEVIS als Datenprojekt
4. Wie geht es weiter?

Freie Universität Berlin

e-Science - Intelligent Openness

Welchen Herausforderungen müssen wir uns stellen, um dem aktuellen Forschungsrahmen Rechnung zu tragen?

→ *Science as an Open Enterprise (Royal Society Report, 2012)*

wichtige Forschungsdateneigenschaften ...

- zugänglich (accessible)
- bewertbar (assessable)
- verständlich (intelligible)
- verwendbar (usable)

Rahmenbedingungen

→ langfristige Sicherung der Forschungsdatenbestände

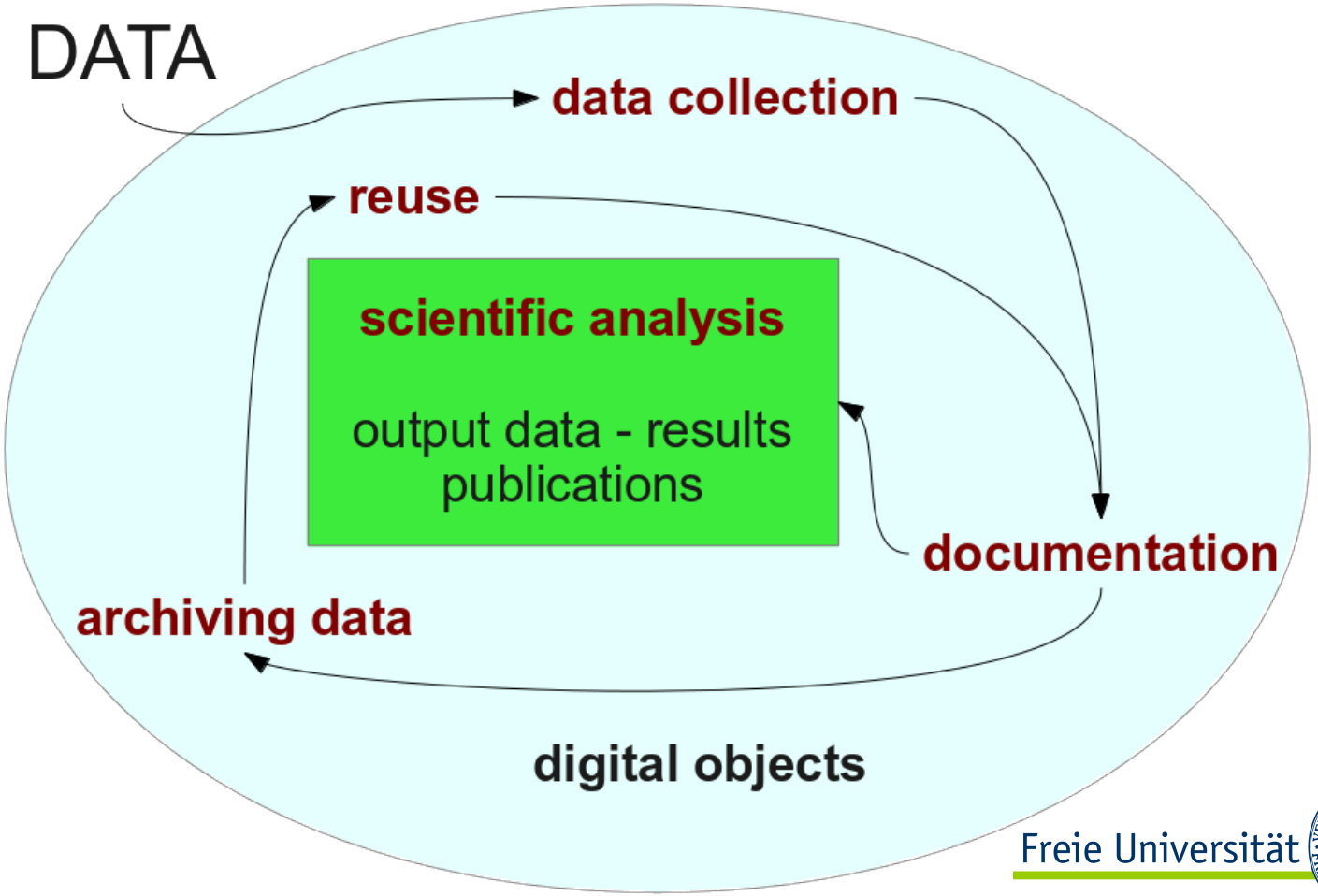
→ (fachübergreifende) Nachnutzung von Forschungsdaten

→ Vernetzung von Datenbeständen

→ Empfehlung der DFG zur 10-jährigen Sicherung von Primärdaten

→ Konzepte: Lebenszyklus, Domänen und Übergänge

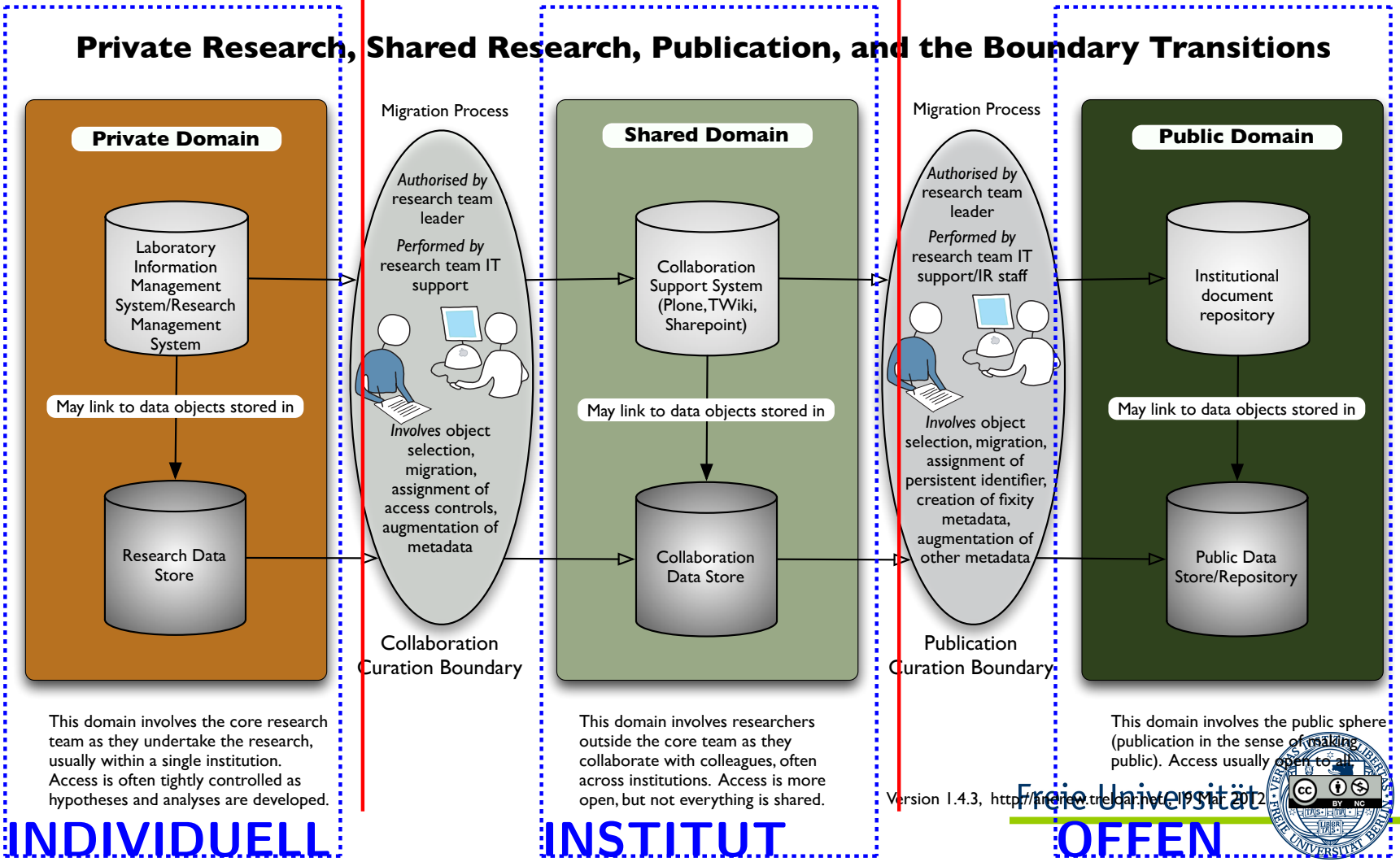
Lebenszyklus von Daten



Freie Universität Berlin

Domänen nach *Treloar (2012)*

→ Kollaboration → Publikation



Version 1.4.3, http://ewig.treloar.net, 19 Mar 2012

Vertiefung

Empfehlung von Quellen ...

Handbuch Forschungsdatenmanagement (Büttner, Hobohm, Müller, Bock+Herchen Verlag, Bad Honnef, 2011)

Wiki http://www.forschungsdaten.org/

Handreichung zum "Einstieg ins Forschungsdatenmanagement in den Geowissenschaften" (2014)

Umsetzung

Was kann auf Institutsebene im Sinne von INTELLIGENT OPENNESS getan werden?

Übersicht zum Projekt EWIG

Entwicklung von **W**orkflowkomponenten für die Langzeitarchivierung von Forschungsdaten **i**n den **G**eowissenschaften

Beteiligte	ZIB, GFZ, IfM
Laufzeit	Nov/2011-Okt/2014
Förderer	DFG - Förderbereich Literaturversorgungs- und InformationsSysteme (LIS)
Ressourcen	0.5 Wissenschaftler pro Einrichtung, SHKs

Freie Universität Berlin

Forschungsfragen

? **unsere TODO Liste**

- Workflowkomponenten und Schnittstellen zur Langzeitarchivierung identifizieren
- Musterlösungen zur Übergabe von Forschungsdaten entwickeln
- Prozess der Langzeitarchivierung von kontinuierlich wachsenden Datensätzen beispielhaft implementieren
- "Wege in die Köpfe" beschreiten

! **Welche Ergebnisse wurden erreicht?**

Ergebnisse aus EWIG

in 3 Jahren Projektarbeit entstand ...

- Policy für Datenprojekte am IfM
  ... und deren Umsetzung in die Praxis
- "Einstieg ins Forschungsdatenmanagement in den Geowissenschaften" (Handreichung für Lehre und Forschung)
  ... eingesetzt in Bachelor Modul "Statistik und Programmierung"
- MEVIS-Datenprojekt und Portal
  ... exemplarische Umsetzung von "intelligent openness"

Policy zu Datenprojekten

Policy über die Organisation der Datenprojekte

(Stand Juli/2013, Version 3)

§1

Dieses Dokument regelt die Verantwortlichkeiten bei der Verwaltung von Datenprojekten (DP) und enthält Empfehlungen zur Organisation der Daten innerhalb eines DP. Änderungen an diesem Dokument müssen dem IT-Aktiv vorgelegt und entschieden werden.

§2

Ein DP versteht sich als eine zusammenhängende Partition innerhalb des zentralen Datenbereiches mit einer eigenen UNIX-Gruppe und einem ausgezeichneten Nutzer, dem Datenprojektadministratoraccount (DPAA). Das DP hat einen eindeutigen und aussagekräftigen Datenprojektnamen (DPN). Der DPN wird ebenso als UNIX-Gruppenname, DPAA und als Ordnername im Verzeichnissystem benutzt. Das DP wird auf Systemebene der IT-Infrastruktur unter dem Pfad „/daten/[DPN]/“ sichtbar gemacht.

§3

Ein neues DP kann eingerichtet werden, sobald eine inhaltliche und technische Abgrenzung von Daten notwendig wird. Wenn der dafür notwendige Bereich mehr als 5% des noch verfügbaren Plattenplatzes oder mehr als 1% des Gesamtbereiches im zentralen Speicherpool beträgt, muss die Einrichtung durch das IT-Aktiv oder den geschäftsführenden Direktor bestätigt werden. Gleiches trifft bei einer Vergrößerung eines bestehenden DP zu. Das IT-Aktiv kann die Einrichtung eines DP empfehlen.

§4

Das Datenprojekt enthält die zwei Unterordner „work“ und „arch“. Der Ordner „work“ dient als zeitweilig zu nutzender Arbeitsbereich für die Vorbereitung und Bearbeitung von Daten, die nach Abschluss der Arbeiten durch den Projektadministrator in den Ordner „arch“ einsortiert werden. Eine Beschreibung des DP erfolgt in der Textdatei „/daten/[DPN]/README“.

§5

Ein DP wird von mindestens einem Wissenschaftler genutzt. Ein Wissenschaftler aus der Nutzergruppe muss als Datenprojektverantwortlicher (DPV) bestimmt werden. Der DPV kann einen oder mehrere Stellvertreter bestimmen. Dies ist in der „README“ des DP zu dokumentieren.

§6

Aufgaben des DPV:

- sorgt für eine optimale Nutzung des DP durch regelmäßige Überprüfung der DP-Aktivitäten
- dokumentiert das DP in einer README-Datei nach den vorgegebenen Empfehlungen (siehe dazu das Template im Intranet http://...)
- organisiert die Daten im „arch“ Bereich und dokumentiert diese DP-spezifischen Regeln
- sorgt für eine konsistente Überleitung von Daten aus dem „work“ in den „arch“ Bereich
- nutzt den DPAA für administrative Arbeiten im DP
- legt die Verfallszeit von Daten im „work“-Bereich fest
- zeigt Änderungen zum DP (Größenanpassungen, Verantwortlichkeiten) in schriftlicher Form der IT-Administration an

§7

Jede Person mit einem Institutsaccount kann die Schreibrechte für ein DP bekommen, wenn dies im Rahmen seiner Tätigkeit notwendig wird. Der DPV des DP muss für einen Institutsaccount die Aufnahme in oder das Entfernen aus der DP Nutzergruppe bestätigen.

§8

Die Nutzung eines DP setzt die Anerkennung dieser Policy voraus. Bei Missachtung der Policy kann die Nutzung des DP durch das IT-Aktiv unterbunden werden. Darüber hinaus gelten die DP-spezifischen Organisationregeln, deren Einhaltung in der Verantwortung des DPV liegt.

Empfehlungen für den DPV und die Nutzer von DPs

- Der DPV hat die Verantwortung für einen sorgfältigen Umgang mit dem DPAA. Es sollten nicht mehr als zwei Personen diesen Account benutzen. Es wird empfohlen, das Passwort des DPAA regelmäßig zu wechseln.
- Zur Überwachung der Platzaufteilung innerhalb eines DP werden einmal täglich die DP ausgewertet und eine Belegungsliste für jeden Nutzer des DP erzeugt. Diese Liste befindet sich im Datenprojekt unter folgendem Namen „/daten/[DPN]/.[USERNAME] und enthält in der ersten Spalte die Dateigröße, in der zweiten den kompletten Namen der Datei. Diese Dateien sind Grundlage der Statistik unter http://....
- Der DPV sollte vor Benutzung eines DP die Organisationregeln des „arch“-Bereiches in der „README“ dokumentieren. Beispiele für eine Gestaltung der Ordnerstruktur befinden sich auf den Webseiten des Intranet (http://...).
- Der DPV kann jederzeit Daten aus dem Datenprojekt löschen. Dies sollte jedoch immer nach Absprache mit den DP-Nutzern erfolgen.

→ *http://www.geo.fu-berlin.de/met/internes/datenprojekte/*

Freie Universität Berlin

Policy

minimales Regelwerk zur Organisation des **institutionellen Repositoriums** am IfM - "Datenpool"

- Verantwortlichkeiten
- Zugriffsrechte
- Gliederung/Struktur der Plattenbereiche

→ wirtschaftliche Verteilung und Verwaltung der IT-Ressourcen
(ca. 280 TB momentan, Wiederbeschaffungswert ca. 100 T€)

Wege in die Köpfe - 3./4. Juli 2014

Workshop zur Entwicklung einer Handreichung für die Lehre

Freie Universität Berlin

MEVIS

Messwert**E**rfassungs-, **V**erarbeitungs- und **I**nformations**S**ystem (Fa. Thies)

... Einbindung in das institutionelle Repositorium
... flexibler Zugang zu den Daten
→ **Erweiterungen ...**

- Datenprojekt und Archivierung "/daten/mevis/arch/"
- Softwarestack "mevis-lib"
- Dokumentation (Sensornetz, Datenportal, Software)
- MEVIS-Portal **http://mevis-www.met.fu-berlin.de/**

Freie Universität Berlin

Arbeitsschritte

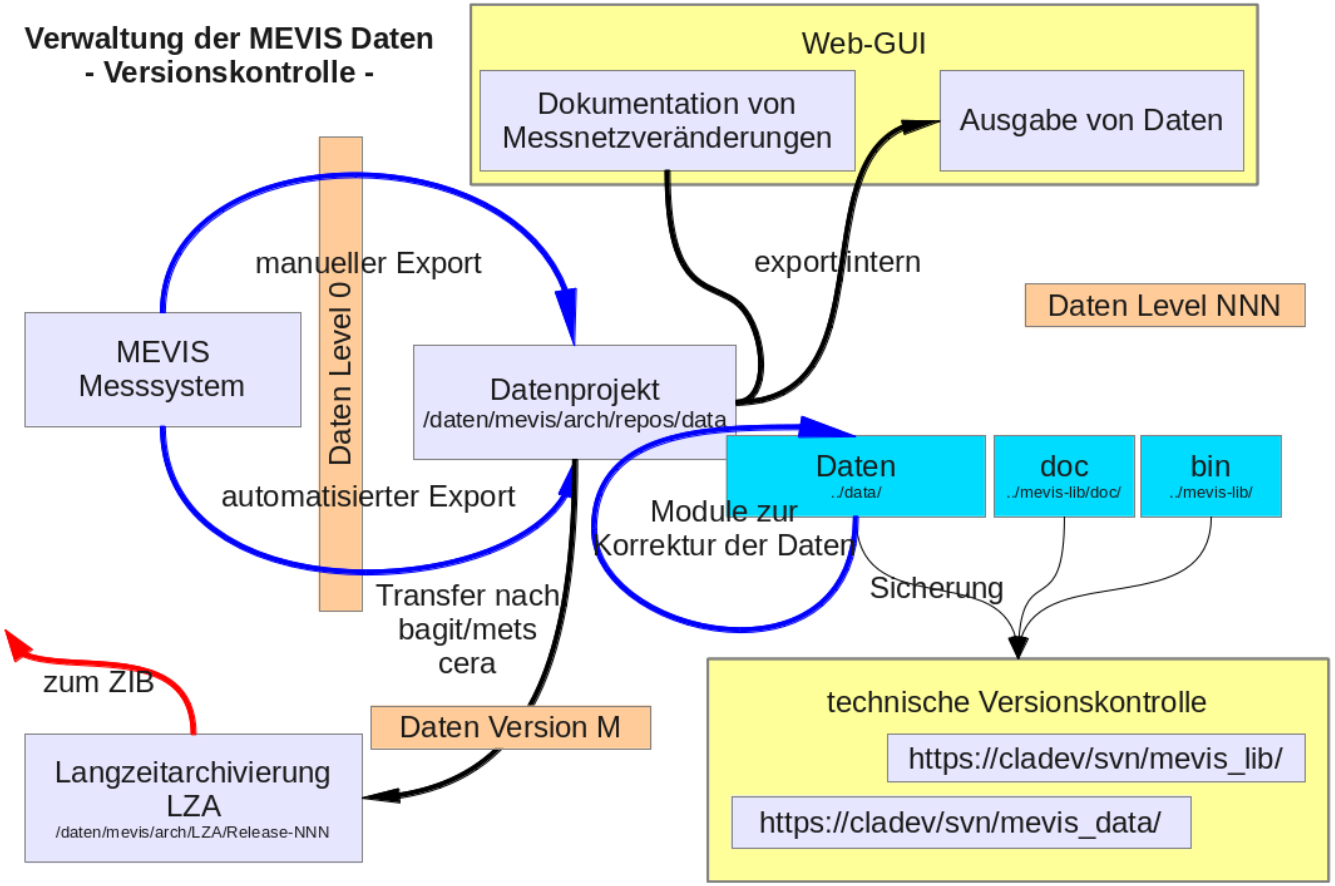
1. Analyse der Ausgangslage
2. Gestaltung des Workflows vom Sensornetz bis zu einer Datenablage, die den Rahmenbedingungen von "intelligent openness" genügt
3. Implementierung der Komponenten des Systems im institutionellen Repositorium

Ausgangslage

- heterogenes Sensornetzwerk
- abgeschlossener Datensammler, kein direkter Zugang zu den Messwerten
- Rohdaten sind Minutenwerte

Aufgabe: Transfer der Rohdaten aus dem Sensorsystem in ein Langzeitarchiv

Datenfluss - MEVIS → /daten/mevis/



Bereitstellung der Daten

- möglichst flexibel
- nicht frei zugänglich
- transparent für die Nutzer
- effizient, mit wenig Personaleinsatz
- reproduzierbare Auslieferung

Lösungsansatz

Die Daten des Sensornetzwerkes können auf zwei Wegen genutzt werden ...

1. auf der Kommandozeile → Python Programmbibliothek
2. Webbasiert → Mevis-Portal

Struktur des Datenprojektes

- **/daten/mevis/arch/repos/data/**
  Minutenwerte nach Station und Sensor sortiert, pro Datei immer ein Tag erfaßt, Datenformat ASCII
- **/daten/mevis/arch/repos/mevis-lib/**
  Dokumentation, Python-Scripte, Qualitätsmodule
- **/daten/mevis/arch/LTA/**
  Zusammenstellung der Daten für den Transfer in ein LTA

→ Ausgabe der Daten als Tabelle (ASCII)
→ Einbindung in Python über die mevis-Klasse

Statistik

Anzahl der Stationen	12 (von 13)
Anzahl der Messstellen	17 (von 19)
Anzahl von Sensoren	104 verschiedene
verschiedene Größen	47
verschiedene Elemente	ca. 20
Zeitraum	Feb/1997 - Jul/2014
Speicherplatz	ca. 130 GB
Dateien	ca. 541000

Hauptstation

Station	Sensoren	Zeitraum	Tage
Steglitz (Feb/1997-Jul/2014)			
Botanischer Garten (1)	11	Feb/1997 - Jul/2014	6392
Botanischer Garten (2)	9	Feb/2008 - Jul/2014	2374
Fichtenberg	16	Feb/1999 - Dez/2001	1066
FB-Messturm (1)	16	Feb/2001 - Dez/2013	4718
FB-Messwiese (1)	6	Feb/2001 - Dez/2010	3622
FB-Messwiese (2)	1	Mär/2010 - Jul/2014	1616
nicht erfaßte Stationen			
Albrecht Thaer Weg			
FB-Messturm (2)			

Freie Universität Berlin

Netzwerk Stationen

Station	Sensoren	Zeitraum	Tage
Steglitz	bis 16	Feb/**1997** - Jul/2014	6392
Gatow	5	Feb/**1997** - Jul/2014	6391
Tempelhof	14	Feb/**1997** - Dez/*2011*	5448
Marzahn	10	Feb/1999 - Mai/2014	5589
Tegel Forstamt	6	Feb/1999 - Jul/2014	5661
Fasanenstrasse	6	Feb/2000 - Jul/2014	5296
Wannsee	9	Feb/2003 - Jul/2014	4201
Planetarium	4	Feb/2005 - Jul/2014	3453
Pichelsdorf	9	Feb/2007 - Jul/2014	2737
Funkturm	6	Feb/2009 - Dez/*2010*	700
Tegeler See	7	Jan/2009 - Jul/2014	2038
Mueggelsee	9	Feb/2011 - Jul/2014	1278

Mevis-Portal

Was bietet die WebGUI?

1. flexible Auswahl der Daten nach Station, Sensor, Jahr
2. Feinauswahl des Zeitraums
3. Zusammenstellung der ausgewählten Daten im Hintergrund
4. Dokumentation der Stationen
5. Anleitung und Dokumentation des Datenprojekts
   → kurze Demonstration
   http://mevis-www.met.fu-berlin.de
6. *Aktualisierung der Historie und Sensorbeschreibung*

Qualitätsmanagement

- Bereitstellung von Python-Modulen in der “mevis-lib”
- Datenablage berücksichtigt Buchführung über Korrekturen einzelner Werte
- Markierung von Qualitätsstufen möglich
- Korrekturprozess muss manuell durchgeführt werden, QM-Module liefern Entscheidungshilfe durch Markierung von zweifelhaften Messungen

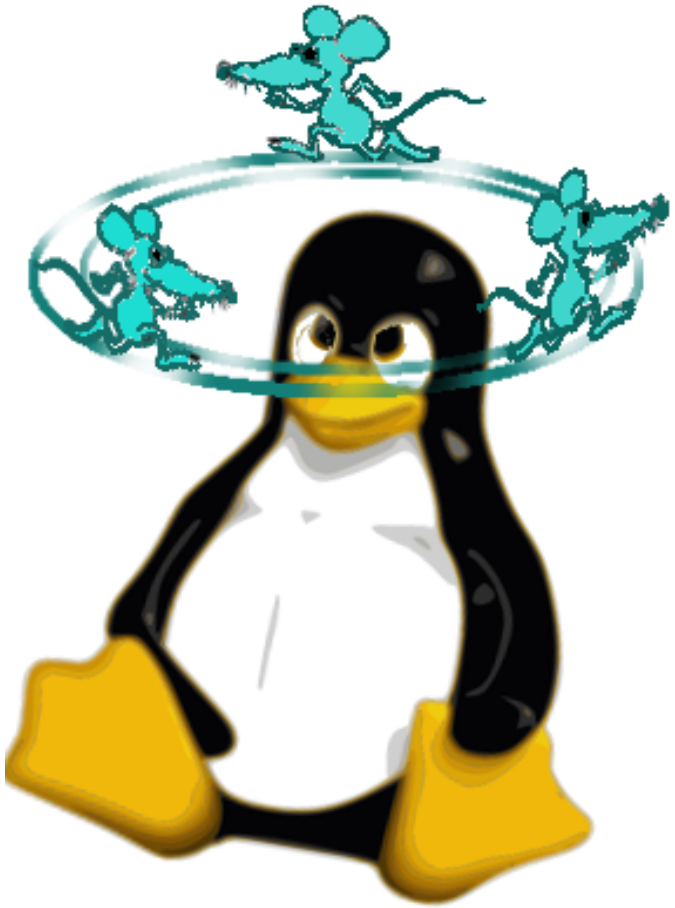
Zur Diskussion steht ...

- Wo besteht Entwicklungspotential?
- Welche Aufgaben sind in Routine zu überführen?
- Wer übernimmt welche Rolle?

Vorschläge

- Anwendung der bestehenden QM-Module auf die wichtigsten meteorologischen Größen aller Stationen
- Erstellung eines geprüften Datensatzes für den Zeitraum 1997-2014
  $\rightarrow$ Datenveröffentlichung
- kontinuierliche Ergänzung der Daten (monatlich)
- Erweiterung der QM-Module
  $\rightarrow$ Bachelor- Masterarbeiten
- Portalerweiterung zur Dokumentation der Messnetzhistorie und der Stationen

Danke für die Aufmerksamkeit



Freie Universität Berlin