



Stiftung Alfred-Wegener-Institut
für Polar- und Meeresforschung
in der Helmholtz-Gemeinschaft



Daten-Archive und die DFG- „Regeln guter wissenschaftlicher Praxis“

*H. Pfeiffenberger, Ch. Wübber
Alfred Wegener Institut , Bremerhaven*

Empfehlung 7 der DFG (1998)

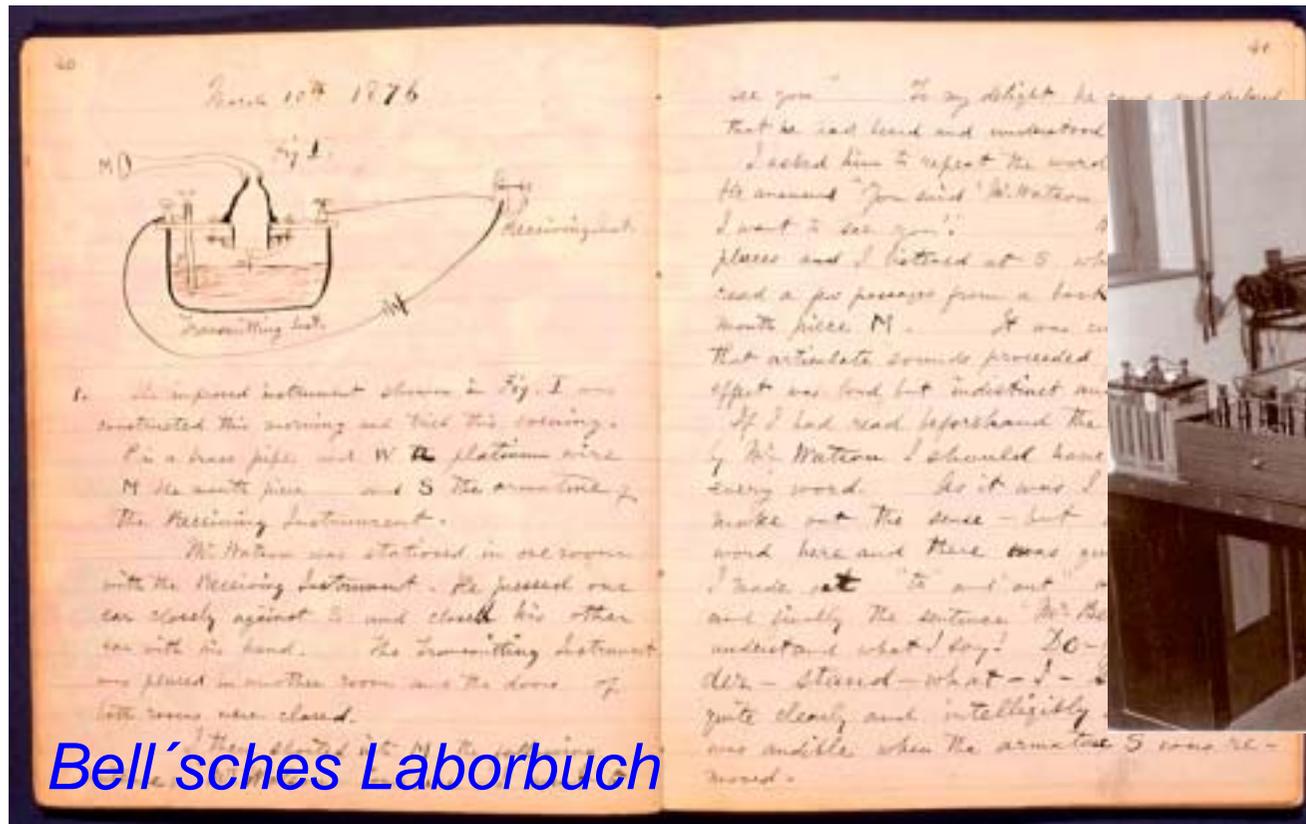


- *Kommission „Selbstkontrolle der Wissenschaft“*
 - **<http://www.dfg.de/antragstellung/>**
- *Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für **zehn Jahre** aufbewahrt werden.*
 - **Auf die Aufzeichnungen später zurückgreifen zu können, ist schon aus Gründen der **Arbeitsökonomie** in einer Gruppe ein zwingendes Gebot. Noch wichtiger wird dies, wenn veröffentlichte Resultate von anderen angezweifelt werden.**
 - **Schon deshalb ist die Feststellung wichtig, dass das Abhandenkommen von Originaldaten aus einem Labor gegen Grundregeln wissenschaftlicher Sorgfalt verstößt und **prima facie** einen Verdacht unredlichen oder grob fahrlässigen Verhaltens rechtfertigt**
 - **Daher hat jedes Forschungsinstitut, in dem lege artis gearbeitet wird, klare Regeln über die **Aufzeichnungen**, die zu führen sind, und über die **Aufbewahrung** der Originaldaten und Datenträger ... zu erlassen.**

Hintergrund - was hat sich geändert?



- Soziale „Kontrolle“ in größerer Community
- Anderer finanzieller Maßstab
- Andere (Daten-) Mengenskala



Bell'sches Laborbuch



Am Helmholtz-Pendel

Hg. I: Soziales und Finanzielles



- *Daten-„Skandale“ in Medizin (98) und Physik (03):*
 - **Anreiz, zu „erfinden“ ist offenbar gewachsen**
 - **Offenbare technische und organisatorische Lücken**
- *Datengrundlagen wurden von den Peers nicht (rechtzeitig) überprüft („Ablauforganisation“)*
 - **nicht vom Vorgesetzten**
 - **nicht von Ko-Autoren**
 - **nicht von den Gutachtern**
- *Letzte Ausrede: „Platte war voll“*
 - **Wären überall die **technischen Mittel** verfügbar?**
 - **Wird die **Zeit gewährt**, die Daten zu beschreiben und zu „formatieren“ und abzulegen? („lege artis“)**

Hg. II: *Finanzielles, Zeit und Menge*



- *Undokumentierte Daten in obskuren Formaten auf obsoleten Datenträgern:
Ein nicht zu hebender Schatz?*
 - **Satelliten-Bilder der NASA: Turnhallen voller 9-Spur-Bänder (70er Jahre?, „Der Spiegel“) => DFD (DLR)**
- *Dokumentierte Daten:*
 - **Elementarteilchen-Beschleuniger (z.B. DESY)**
- Ein *Beispiel vom AWI: See-Seismik*
 - **ca. 12 TB aus 20 Jahren**
 - **Standardformate, -Software**
 - **entsprechen Jahren an Schiffsnutzung => 100 M€**
 - **Kopieren von 25.000 Bänder, Spiegeln (Neubaumittel)**

H. III: Die (Selbst-) Organisation wissenschaftlicher Arbeit



- *Hauptmittel wissenschaftlicher Qualitätssicherung: **Nachvollziehbarkeit** im Labor, auf dem „Aktenwege“*
- *Historische „Aktenlage“:
Laborbuch, Veröffentlichungen, Vorträge*
- *Heute:*
 - **Nachvollziehen im Labor: Unrealistisch => Kosten !**
 - **Großteil der „Aktenlage“ : Datenträger (?) => Auffindbar ?**
- *„Laborbuch“ und (klassische) Publikation reichen nicht!*
- ***Aber : Gute Daten-Praxis ist de facto (noch) nicht (durchgehend) definiert und etabliert!***

Konsequenzen für RZs (Bottom Up)



- *Datenmanagement* => *Informationsmanagement*
- *Desaster-Recovery* => *Archivierung*

- *Accountmanagement* => *Identitätsmanagement*
- *Datenintegrität* => *Non-Repudiability*

- *Organisation der Requirement-Analyse*
 - **Angemessenheit**
 - **Best Practise (je Community)**
 - **Abläufe**

Informationsmanagement



- *(Meta-)Datenbanken* - hier nicht i.S. verteilter Dateisysteme, sondern Suchkriterien
- Von dort : *Persistente "Links"* zu Flat Files
- Einführen und Verwalten von *DOIs* oder ähnlichem

- Schliesslich : *(Bidirektionale) Persistente Links* zwischen

Publikationen, Daten, Metadaten und Personen

- *Workflows* zum Sicherstellen der Verbindungen

Archivierung



- *Desastervorsorge: Nicht (nur) Backup, sondern auch **Spiegelung (anderes Gebäude)***
- *Davon eine Kopie **WORM** ? (=> kaufmännische Sitte)*
- *Datenträger-Management:*
 - **Aus dem Schrank des Wissenschaftlers in das HSM**
 - **Regelmäßige Auffrischung**
- *Formatmanagement (Daten und Metadaten)*
 - **technische Formate (XML)**
 - **Semantik (Dublin Core)**
- *Change-Management obsoleter Datenträger und Datenformate (Bsp.: Polarsterndaten: 20a, 3 Systeme)*

Identity Management



- Workflows bedingen : **Identität, Berechtigungen, Rollen** => mehr Anforderungen als für Accounting
- Identitäten müssen ebenfalls mindestens 10 a aufrecht erhalten werden
- danach zumindest **kein Recycling von IDs**
 - Alternativ: „Anonymisieren“ von Daten
 - aber: Autorenschaft von Datenpublikationen
- *eduPersonPrincipalName (persistent!), eduPersonAffiliation*
=> *www.Internet2.edu/Middleware => GRID*
- **Föderierte Verzeichnisse (wer darf 1 TB abholen?)**

Integrität, Non-Repudiability



■ *(Arbeits-) Ökonomie*

- **gegen Verlust sichern (Kosten, Wert der Daten)**
- **verfügbar machen und halten**
 - *Ab wann ? (Publizitätspflicht?, Bsp.: NDSC)*
 - *Für wen ?*

=> Förderiertes Identitäts- und Rollenmanagement

■ *wiss. Qualitäts-Sicherung*

- **Integrität sichern**
 - „**unverändert seit Publikation**“
- **Verantwortlichkeit festhalten**
 - „**wer hat als letzte(r) die Daten geändert**“

Requirements-Analyse



- „*tall order*“ => *Angemessenheit, zu messen an:*
 - Wert der Daten
 - „Kosten“ des Verlustes der Glaubwürdigkeit
- *abstrakte Best Practices erfassen*
 - verschieden, je Community
 - **Policies ableiten => Durchsetzen ist Leitungsfunktion**
- *Datenwelten analysieren*
 - Systeme - Flat Files, RDBMS, CVS, ...
 - Formate
 - Metadaten
- *Matrix : Communities / Datenwelten*

Minimal-Konsens Requirements



- *Wozu Konsens?*
 - **Vergleichbarkeit der „Best Practice“**
 - **Interdisziplinäre Zusammenarbeit**
 - **Globale Zusammenarbeit**
- *Daher: Standards für das Minimum*
 - **Metadaten (-formate)**
 - *Dublin Core*
 - *FGDC (georeferenzierte Daten)*
 - **Daten**
 - *Einheiten*
 - *Formate (netCDF)*
 - **Personen (eduPerson)**
 - **Linking (DOIs)**

Beispiele (AWI) I



- *Erdsystem-Modellierung - was archivieren?*
 - Initialisierungsdaten
 - Code (CVS?)
 - Parameter, Jobskripte, Logfiles?
 - Ergebnisse ?? (Rekonstruierbarkeit?)
 - das Ganze noch einmal für Postprocessing ?????
- *CryoSat Bilder - Verantwortlichkeit*
 - erste (integere?) Kopie beim DFD
 - Benachrichtigung bei Korrektur ?
 - muss also doch die „Arbeitskopie“ gesichert werden ?
 - reicht ein Link zur DFD-Kopie ?
 - wer sichert Persistenz ?

Beispiele (AWI) II



*Beispiele neben den „üblichen Verdächtigen“
(Modellierung, HEP und Satelliten) :*

■ *Mikroskopie und Bildwandler*

- **30 MB / Bild ; 10 Bilder / Std, Mikroskop ; 200 Tage / Jahr
=> 0,5 TB je Mikroskop,Jahr
=> 15.000 Metadaten-Einträge je Mikroskop,Jahr**
- **10 Mikroskope, 10 Jahre
=> 50 TB
=> 1,5 Mio Metadaten-Einträge**
- **eine „Datenwelt“, mehrere Communities (Bio, Geo)**

■ *Bioinformatik (Meta-Gendatenbank ALGINET)*

Zusammenfassung I



- *RZs müssen einige Themen ganz neu „sehen“*
 - **(Sinn des) Backup, Spiegelns**
 - **Platten „aufräumen“, reorganisieren (Mountpoints)**
 - **„Benutzer löschen“**
- *Transparenz, Verständlichkeit, Einfachheit, Konsistenz, Konstanz*
 - **keine Entschuldigung : „ich wusste nicht dass...“**
- *neues Maß an „Veröffentlichung“ der Daten*
 - **wird nicht bei allen Nutzern auf Gegenliebe stoßen**
=> kontrollierter Zugang zu Daten (Peers, Gutachter)
=> zeitabhängig (!), Policy-gesteuert

Zusammenfassung II



- *Weite des Horizonts - räumlich und zeitlich*
 - **10 Jahre +**
 - **Persistenz von Links zwischen Objekten**
 - => auf verschiedenen DV-Systemen
 - => in potentiell global verteilter Verantwortlichkeit

- *Semantische , inhaltliche Kompetenz notwendig*
 - => *Kooperation mit Bibliotheken ??*
 - => *was sind die „Kernkompetenzen“ des RZ ?*

- *Verantwortung für den Ruf der Wissenschaftler und der Institution*
 - => *Bedeutung der Rolle des CIO*