Forschungsdatenmanagement in der Genomforschung

Ergebnisbericht

zum

LABIMI/F-Workshop 2012 in Kiel

Romanus Grütz

4. Mai 2012

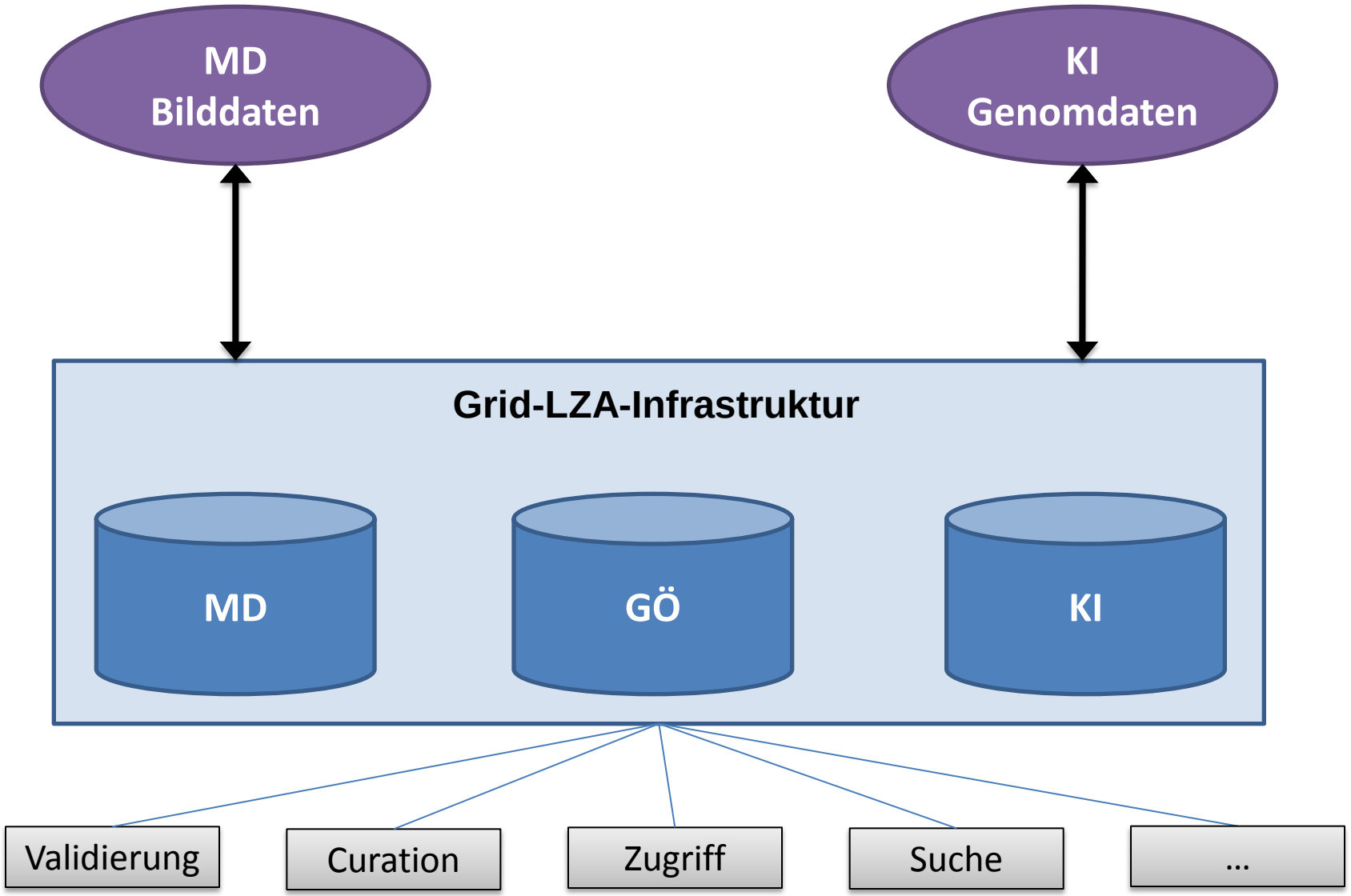
Agenda

1. Vorstellung des Projekts LABIMI/F
2. Ebenen der Langzeitarchivierung
3. Übersicht über den Workshop
4. Zusammenfassung der Ergebnisse
5. Ausblick

Langzeitarchivierung biomed. Forschungsdaten

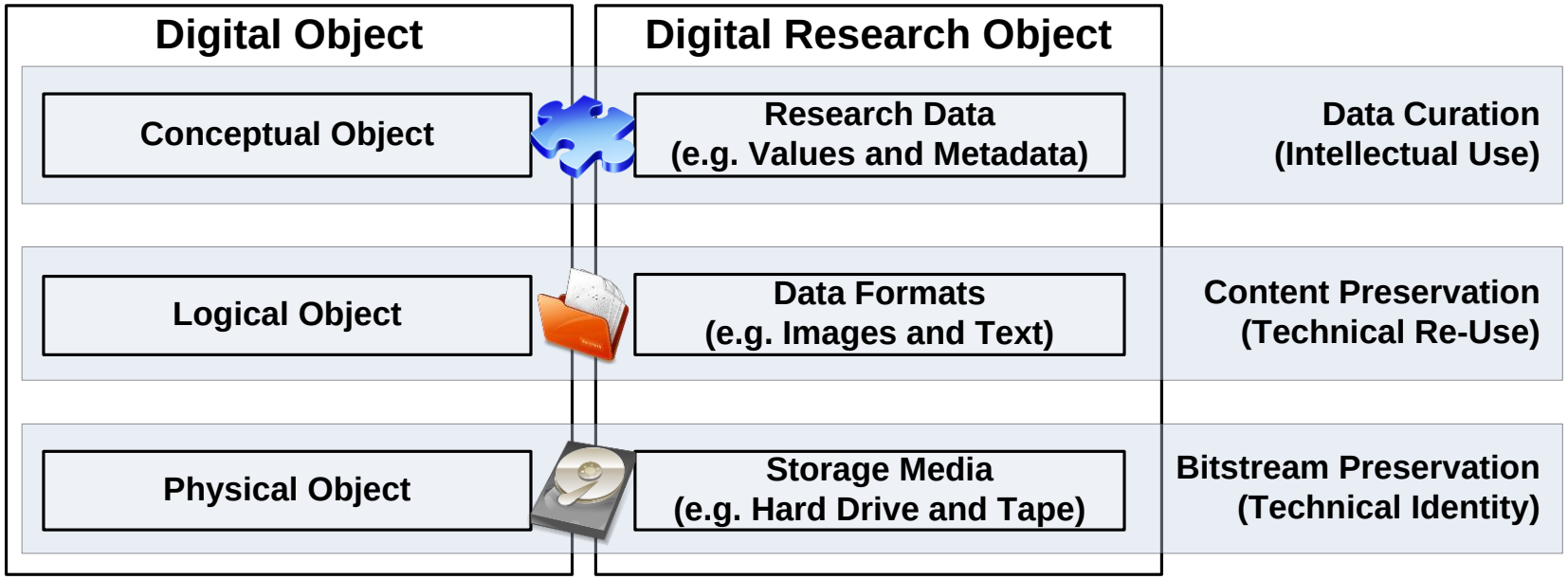
- Gefördert durch die DFG
  - 1998: Vorschläge zur Sicherung guter wissenschaftlicher Praxis
  - 2009: Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten
  - 2010: Antrag in Ausschreibung „Informationsinfrastrukturen für Forschungsdaten“
  - Förderungszeitraum: 06/2011 – 05/2013
- Ziele
  - Spezialisierung auf zwei Anwendungsfälle: Bild- und Genomdaten
  - Erhebung von Anforderungen an Forschungsdatenmanagement aus der Biomedizin
  - Prototypische Implementierung einer Langzeitarchivierungsinfrastruktur auf Basis von Grid Computing

Angestrebtes Ergebnis



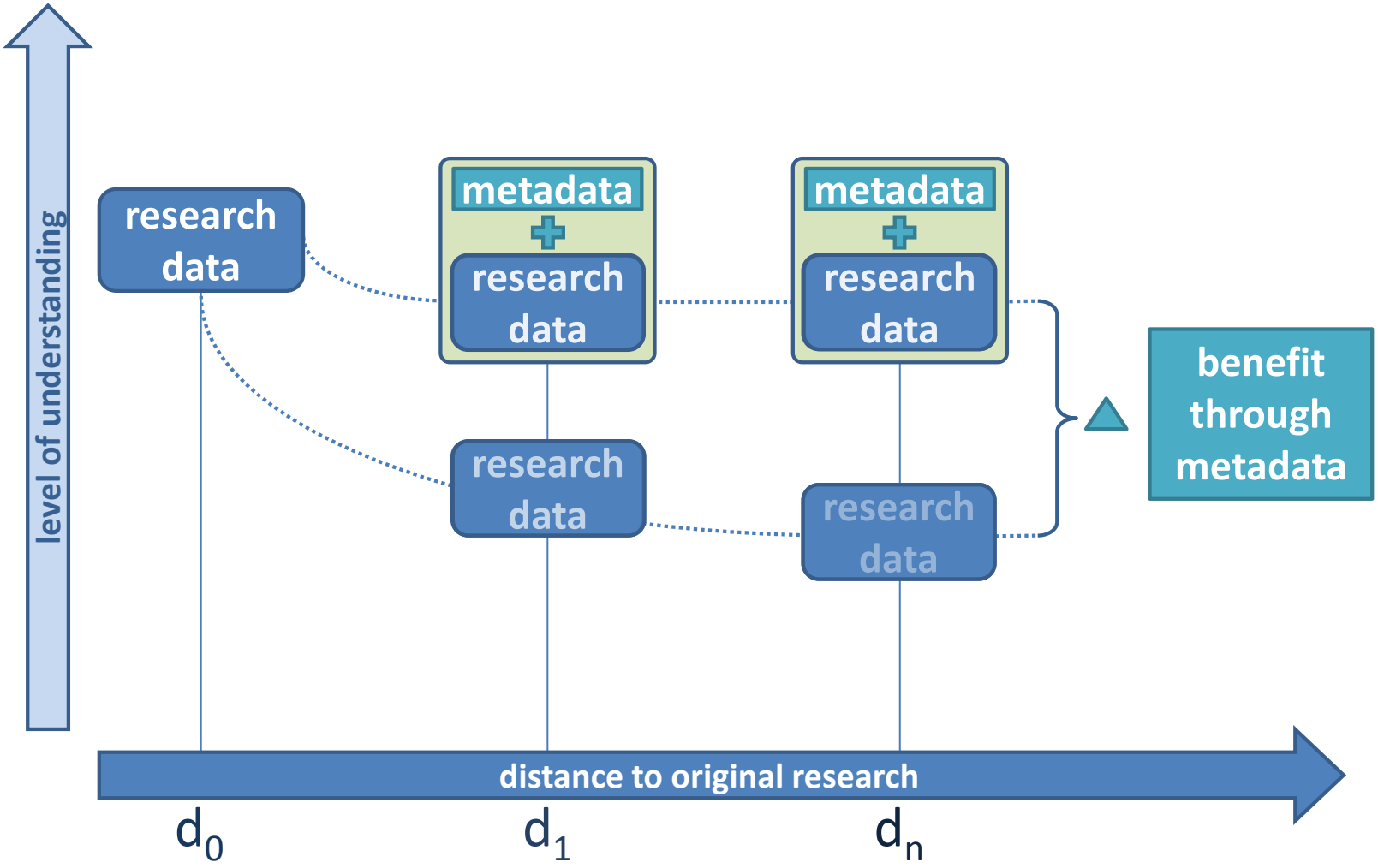
Quelle: F.Dickmann: LABIMI/F - Langzeitarchivierung biomedizinischer Forschungsdaten, http://www.labimi-f.med.uni-goettingen.de/Workshop-Genomforschung/Dickmann.pdf, Abgerufen am 3.5.2012

2. Die Ebenen der Langzeitarchivierung



Modifiziert übernommen von Ludwig, J., Long-term Preservation of Digital Research Data. 2009, DESY: Hamburg, Germany. URL: http://www.desy.de/dvsem/WS0910/ludwig_talk.pdf.

Benefit für Forscher



Quelle: Dickmann, Frank and Grütz, Romanus (2011): LABIMI/F - Digital Preservation of Biomedical Research Data, Knowledge Exchange: Workshop Research Data Management – Activities and Challenges, 14.-15. November 2011, Bonn, Poster, Access date: 2012.01.12, URL: http://www.labimi-f.med.uni-goettingen.de/Publications/Poster_LABIMI-F_v6.pdf..

3. Übersicht über den Workshop

27.03.2012 in Kiel

Thema	Dozent
LABIMI/F - Langzeitarchivierung biomedizinischer Forschungsdaten	F. Dickmann, UMG
Herausforderung: Archivierung genomischer Hochdurchsatzdaten	J. Hampe, UKSH[1]
Aufklärung der Genetik von Hämochromatose und Gallensteinleiden durch "Nachnutzung" von GWAS-Daten	S. Buch, UKSH
Archivierung von NGS-Daten - Artefaktsammlung oder Datenschatz?	M. Nothnagel, UKSH
Datenstandards bei Sequenzierungsdaten	A. Herrmann, UKSH
Efficient data structures for storage and retrieval of multiple biosequences	S. Kurtz, Uni Hamburg
Forschungsdatenmanagement biomedizinischer Genomdaten	M. Wittig, IKMB[2]

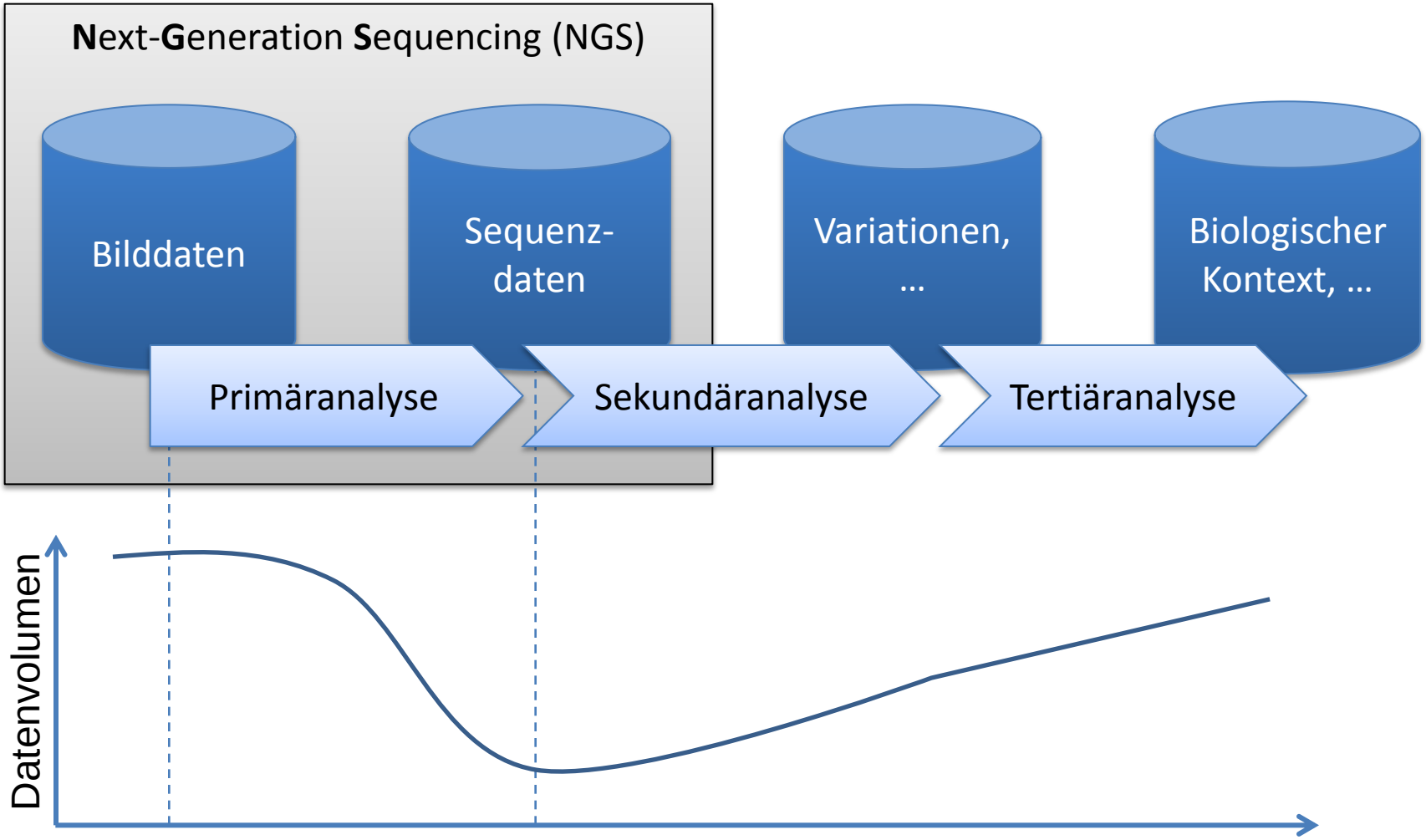
[1] Universitätsklinikum Schleswig-Holstein
[2] Institut für Klinische Molekularbiologie

Zweck / Ausrichtung der Archivierung

- Beweisbarkeit / Forensik
  - Unveränderbarer Audit Trail
  - Strenge Zugriffssteuerung (AAI)
  - Sichere Infrastruktur

- Nachnutzung durch Dritte / Data Sharing
  - Vergabe feingranularer Zugriffsrechte
  - Möglichst hohe (Daten-)Verfügbarkeit / Ausfallsicher
  - Funktionen zum Auffinden von Daten
    - Metadaten notwendig zur Beschreibung / Suche
    - Welche Daten dürfen von wem gefunden werden?

Ablauf



Modifiziert und erweitert übernommen von M. Wittig: Forschungsdatenmanagement biomedizinischer Genomdaten, http://www.labimi-f.med.uni-goettingen.de/Workshop-Genomforschung/Wittig.pdf, Abgerufen am 3.5.2012

Was wird archiviert?

- DNA
  - Biomaterial / Biobank
  - Material ist endlich
  - Verschiedene Sequenzierungsgeräte werden eingesetzt
- Roh-/Bilddaten
  - Hohes Datenvolumen (im 2-stelligen TB-Bereich)
  - Keine Reproduzierbarkeit der Bilddaten möglich
  - Andere Analysen können verwendet werden
- Ergebnisdaten
  - Geringeres Datenvolumen (2 bis 3-stelliger GB-Bereich)
  - Keine Reproduzierbarkeit (ohne Biomaterial) möglich
  - Keine weiteren Analysen (ohne Biomaterial) möglich

Qualität von NGS-Output

- Es gibt drei große NGS-Geräte-Hersteller
  - Applied Biosystems (ABI)
  - Illumina
  - Roche
- Die Sequenzierer verwenden unterschiedliche Techniken und haben unterschiedliche Fehlerprofile

  →Welcher Sequenzierer die geringste Fehlerrate liefert hängt vom Anwendungsfall ab
- Werkzeug zur Qualitätskontrolle von NGS-Rohdaten

  →FastQC, entwickelt vom Babraham Institute

Formate für Genomdaten

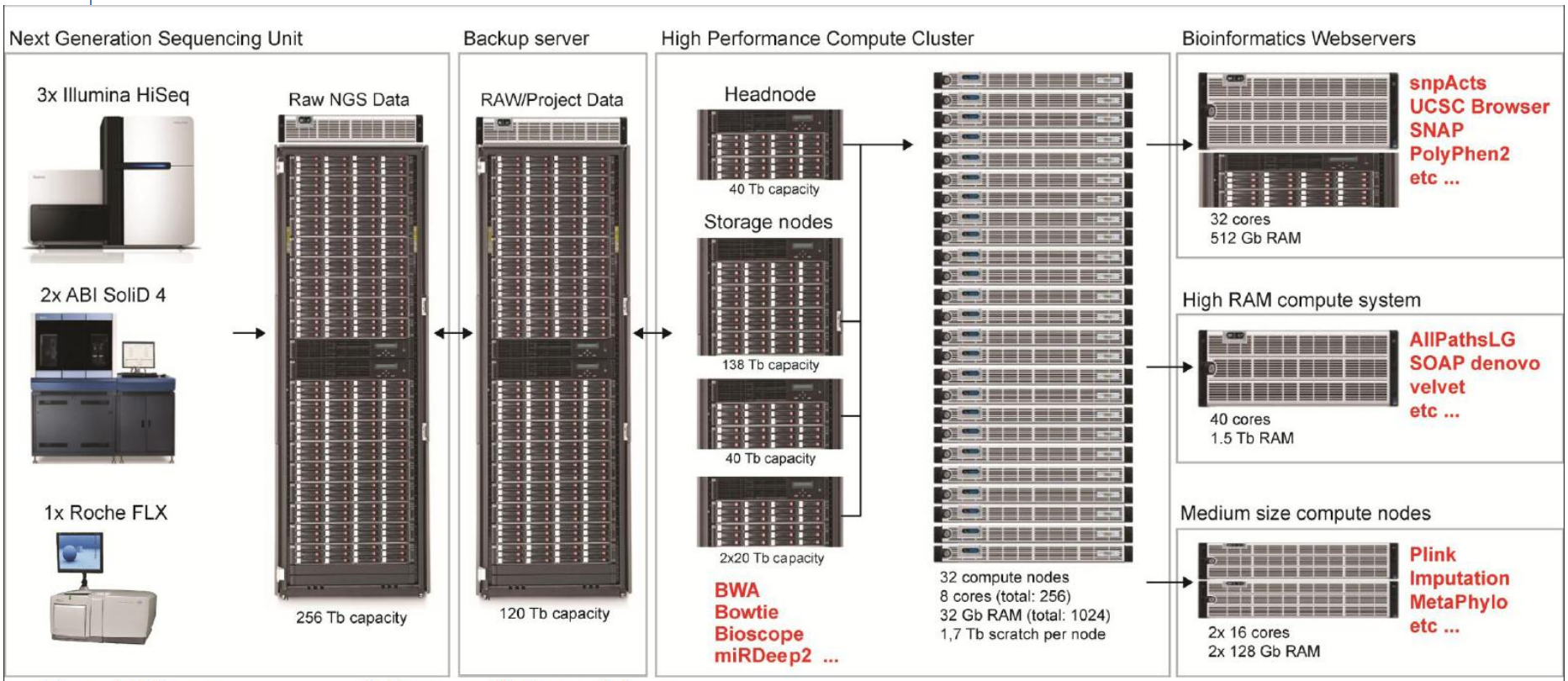
- Die Sequenzierer liefern i.d.R. proprietäre Formate

  →SRA, SRF, SFF, SOLiD_native, llumina_native, PacBio_HDF5, CompleteGenomics_native

- Ergebnisse in Standardformaten
  - FastQ: Sequenzdaten als ASCII: ACTG-Sequenz + Qualitätsinformationen
  - SAM/BAM: Sequence Alignment/Map wird verwendet, um die Daten mit Referenzdaten zu vergleichen
- Es gibt Konverter für die unterschiedlichen Formate

  →Muss noch im Detail betrachtet werden

NGS-Infrastruktur am IKMB



Bsp HiSeq:

2 flow cells/machine
1 flow cell run/10 days
160.000.000 Paired reads/lane (8 lanes/flow cell)
256.000.000.000 bases/flow cell in 10 days
1.536.000.000.000 bases/10 days

Quelle: M.Wittig: Forschungsdatenmanagement biomedizinischer Genomdaten, http://www.labimi-f.med.uni-goettingen.de/Workshop-Genomforschung/Wittig.pdf, Abgerufen am 3.5.2012

5. Ausblick

- NGS ist noch sehr jung
  (im Vergleich zu med. Bilddatenverarbeitung: z.B. PACS)

- NGS-Daten sind fehlerbehaftet und benötigen Qualitätsmanagement; weitere Erfahrungen dazu sind notwendig

- Welche Daten aufgehoben werden sollen, ist noch ungeklärt → hier besteht noch Forschungsbedarf

- Workshop am 25. Juni in Berlin zum Erfahrungsaustausch zu Langzeitarchivierung mit anderen Disziplinen und einem privatwirtschaftlichem Betreiber

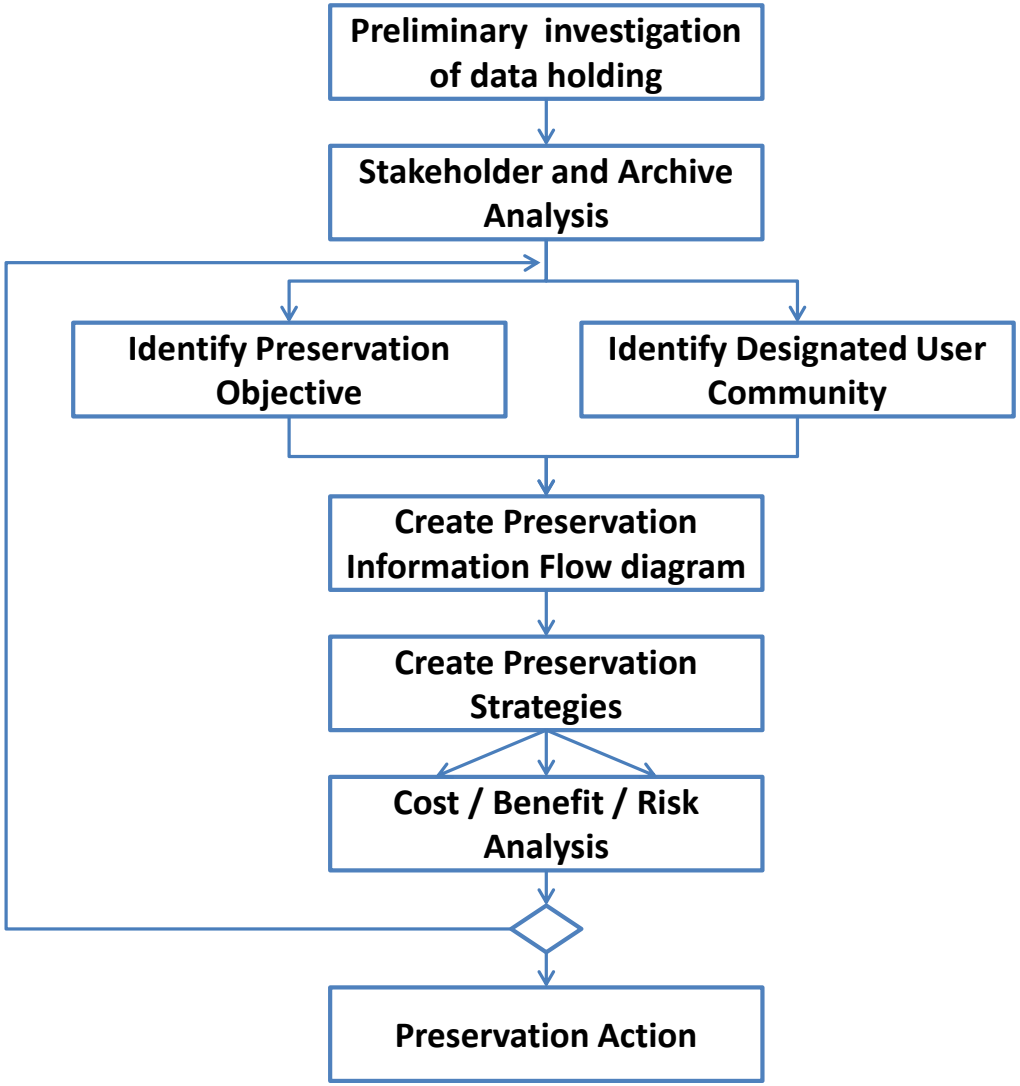
Kontakt

Universitätsmedizin Göttingen
Medizinische Informatik
http://www.mi.med.uni-goettingen.de/

Romanus Grütz

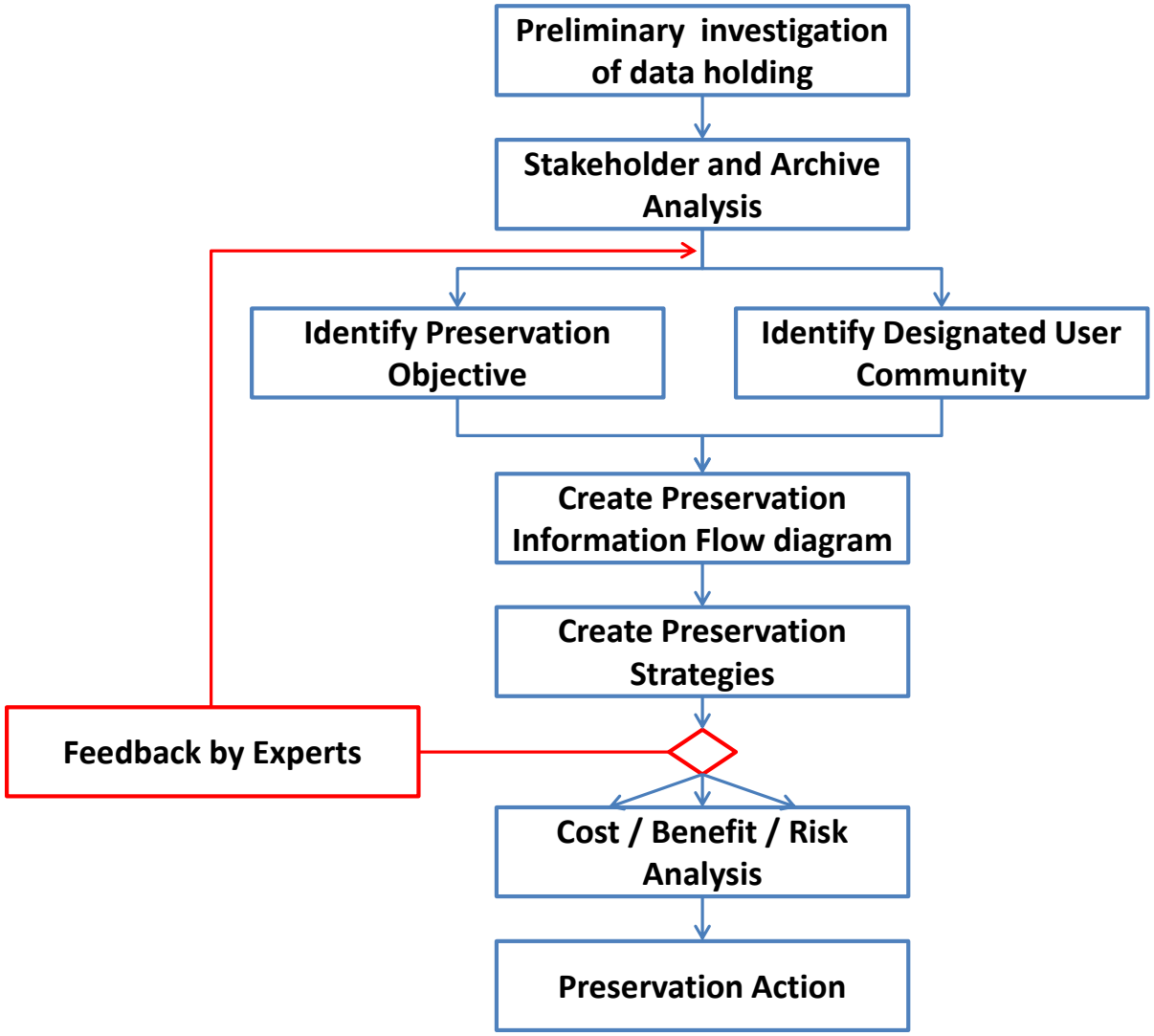
romanus.gruetz@med.uni-goettingen.de
Tel.: (0551) 39 - 6981

Generisches Vorgehenskonzept



Quelle: Conway, Esther; Giaretta, David; Lambert, Simon and Matthews, Brian (2011): Curating Scientific Research Data for the Long Term: A Preservation Analysis Method in Context, International Journal of Digital Curation, Bath, GB, Digital Curation Centre, 6, 2, 2012.01.25, URL: http://www.ijdc.net/index.php/ijdc/article/view/182/264.

Angepasstes Vorgehenskonzept



Adaptiert von: Conway, Esther; Giaretta, David; Lambert, Simon and Matthews, Brian (2011): Curating Scientific Research Data for the Long Term: A Preservation Analysis Method in Context, International Journal of Digital Curation, Bath, GB, Digital Curation Centre, 6, 2, 2012.01.25, URL: http://www.ijdc.net/index.php/ijdc/article/view/182/264.