Infrastruktureinrichtungen und Forschungsdaten: Sichtweise und Services

Jens Ludwig
ludwig@sub.uni-goettingen.de
25. Juni 2012, Berlin

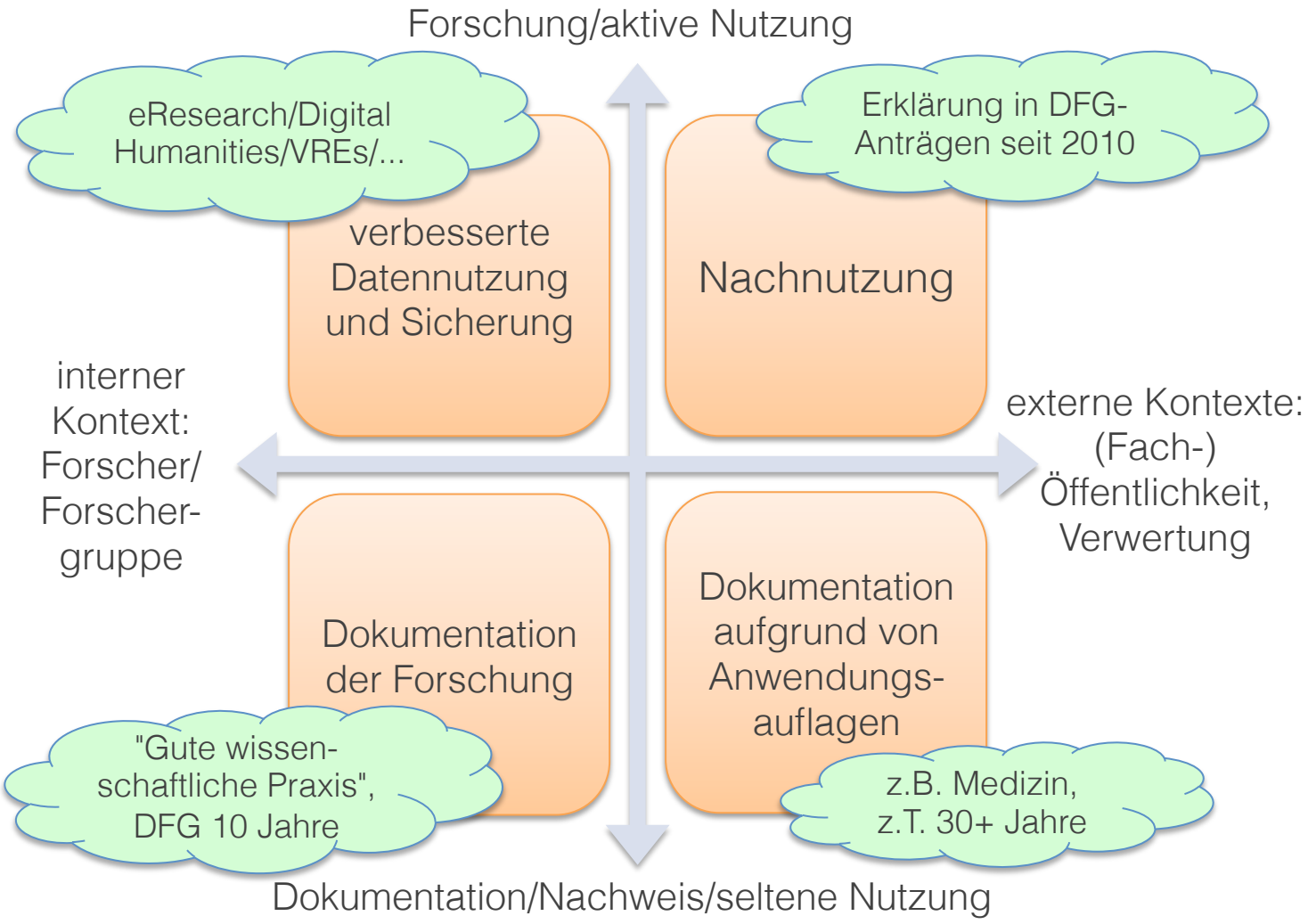
NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN | SUB

Hintergrund

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

NIEDERSÄCHSISCHE STAATS- UND
UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN | SUB

Wozu Forschungsdaten-Management?

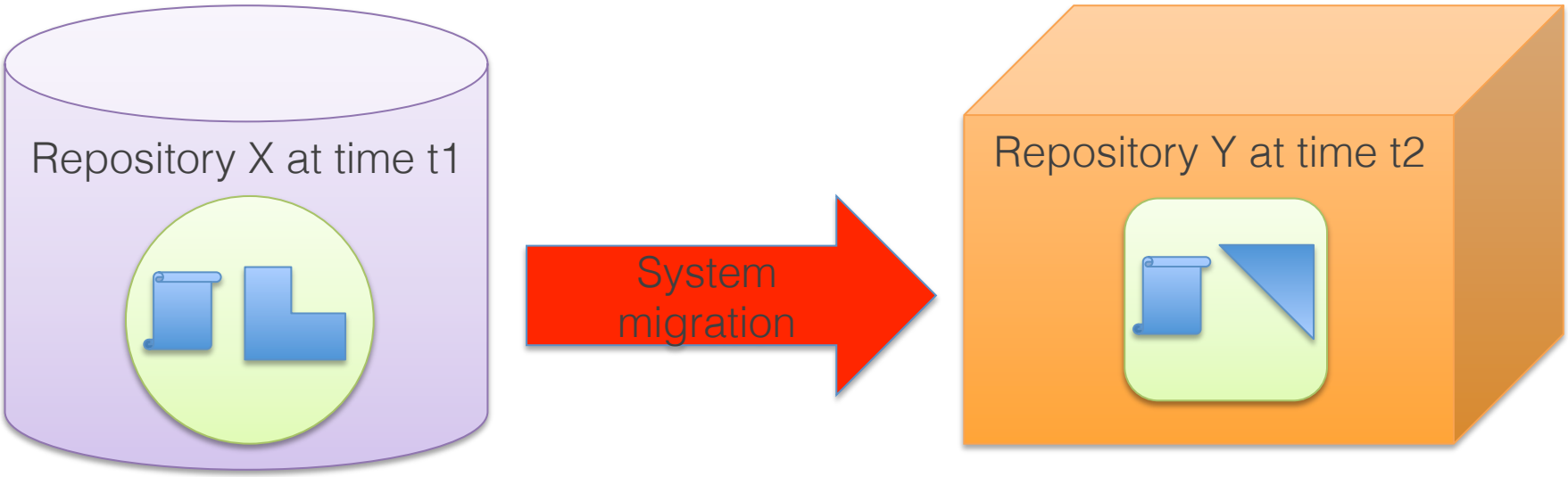


Was ist Langzeitarchivierung (und Forschungsdaten-Management) nicht?

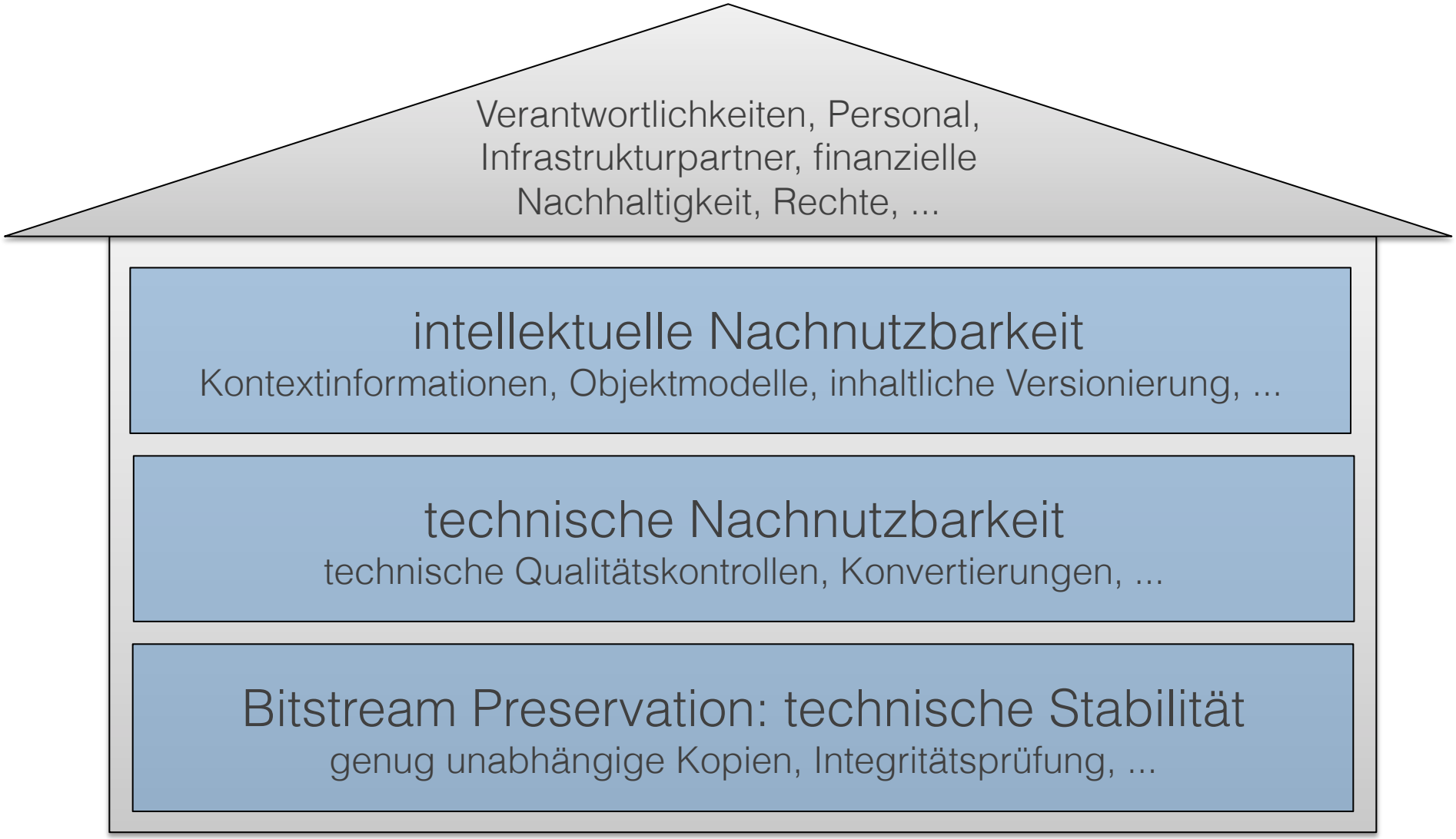
"Preservation is not a Place" (Stephen Abrams et al., California Digital Library)

Auch Archivsysteme veralten und nicht immer ist ein dediziertes Archivsystem sinnvoll.

Nicht technische Systeme, sondern stabile Prozesse (Wartbarkeit und Migrierbarkeit der Infrastruktur), Richtlinien und Organisationen sind die Antwort.

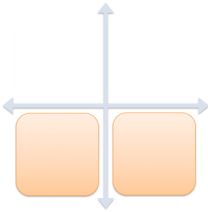


Service Level des Forschungsdaten-Managements



Service-Gruppe: Dokumentation

Ziel: Nachvollziehbarkeit für Verantwortungszwecke



Zielgruppe: Institutionen (weniger Wissenschaftler)

Daten werden sehr selten angefragt

- einfacher und schneller Zugriff wird Kostenersparnis geopfert
- ggf. reicht Bitstream Preservation + Hard-/Software-Museum

Kosten für Dokumentation werden mit Kosten des Verantwortungsfalls abgewogen

Service-Gruppe: Nachnutzung

Ziele:

- zitierfähige Datenpublikation
- erneute wissenschaftliche Nutzung von Daten
- Förderersicht: erhöhte Effizienz
- Bewahrung nicht reproduzierbarer Daten

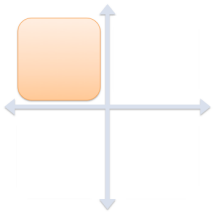
Zielgruppe: Fach-Communities

Daten werden selten benutzt, aber ohne klares Enddatum

Service-Gruppe: verbesserte Datennutzung

Ziele:

- Erleichterung und Absicherung der Datennutzung
- Ermöglichung neuer Methoden/Funktionen



Zielgruppe: wissenschaftliche Arbeitsgruppen, z.B. SFBs

Daten werden ständig benutzt und verändern sich, durch Projektlaufzeit definiert

Verknüpfung mit kollaborativen Forschungsumgebungen und Werkzeugen

Notwendigkeit von Infrastruktur

Komplexe Instrumente können nicht mehr effizient einzeln betrieben werden (z.B. LHC, Teleskope, ...)

Aufwand der Datenerhebung nicht mehr einzeln bewältigbar (z.B. große Befragungen)

Verteilter Untersuchungsgegenstand erfordert Kooperation (z.B. Klimaforschung)

Interdisziplinarität oder Fachdifferenzierung verlangt Kooperation (z.B. Biodiverstität)

Modelle der Dateninfrastruktur

Datenmanagement beinhaltet sowohl fachspezifische als auch generische Aufgaben

Modelle:

- disziplinspezifische Datenzentren (z.B. GESIS, DKRZ)
- disziplinspezifische Föderationen (z.B. LHC)
- disziplinübergreifende Dateninfrastruktur-Einrichtungen (z.B. TIB in DataCite, Rechenzentren, ...)
- institutionsspezifische Lösungen

Welche Infrastruktur und wie disziplinspezifisch muss sie sein?

Einrichtung mit Fachkenntnissen für disziplinspezifische Aufgaben notwendig, übergreifende Einrichtungen aus Effizienzgründen notwendig

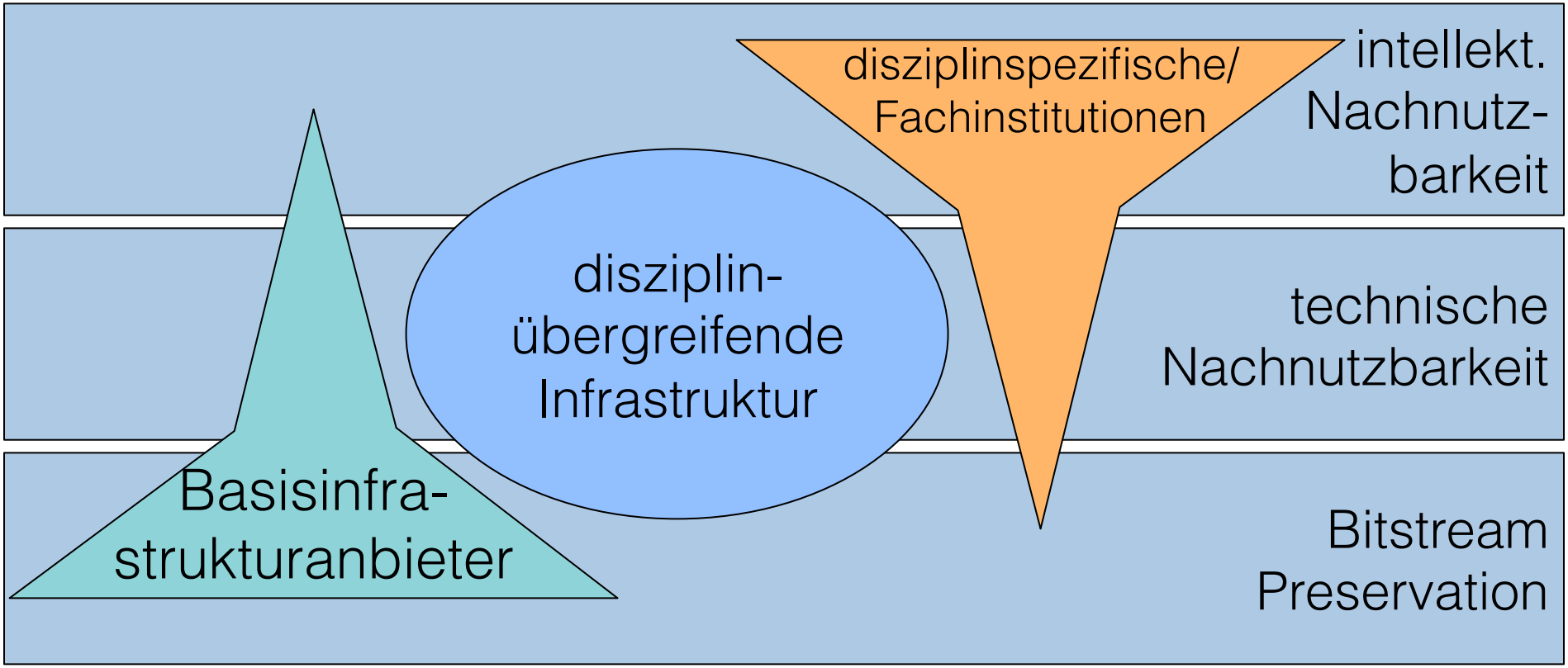
Schwierigkeit der Disziplindefinition: Physik? Astrophysik? Radioastronomie? ...

Datenmanagement ist nie völlig disziplinspezifisch

- Datenmanagement ist eher methodenspezifisch
- Methodenvielfalt in Disziplinen
- erfolgreiche Methoden verbreiten sich in andere Disziplinen (z.B. MRT in Psycholinguistik)

Offene Frage: Welcher Zuschnitt von Infrastruktur ist sinnvoll? Welcher Fachgranularität, welche Fachkenntnissen und welche Aufgaben?

Dienstleister je Service Level



RIN/JISC, "Data centres: their use, value and impact", 2011

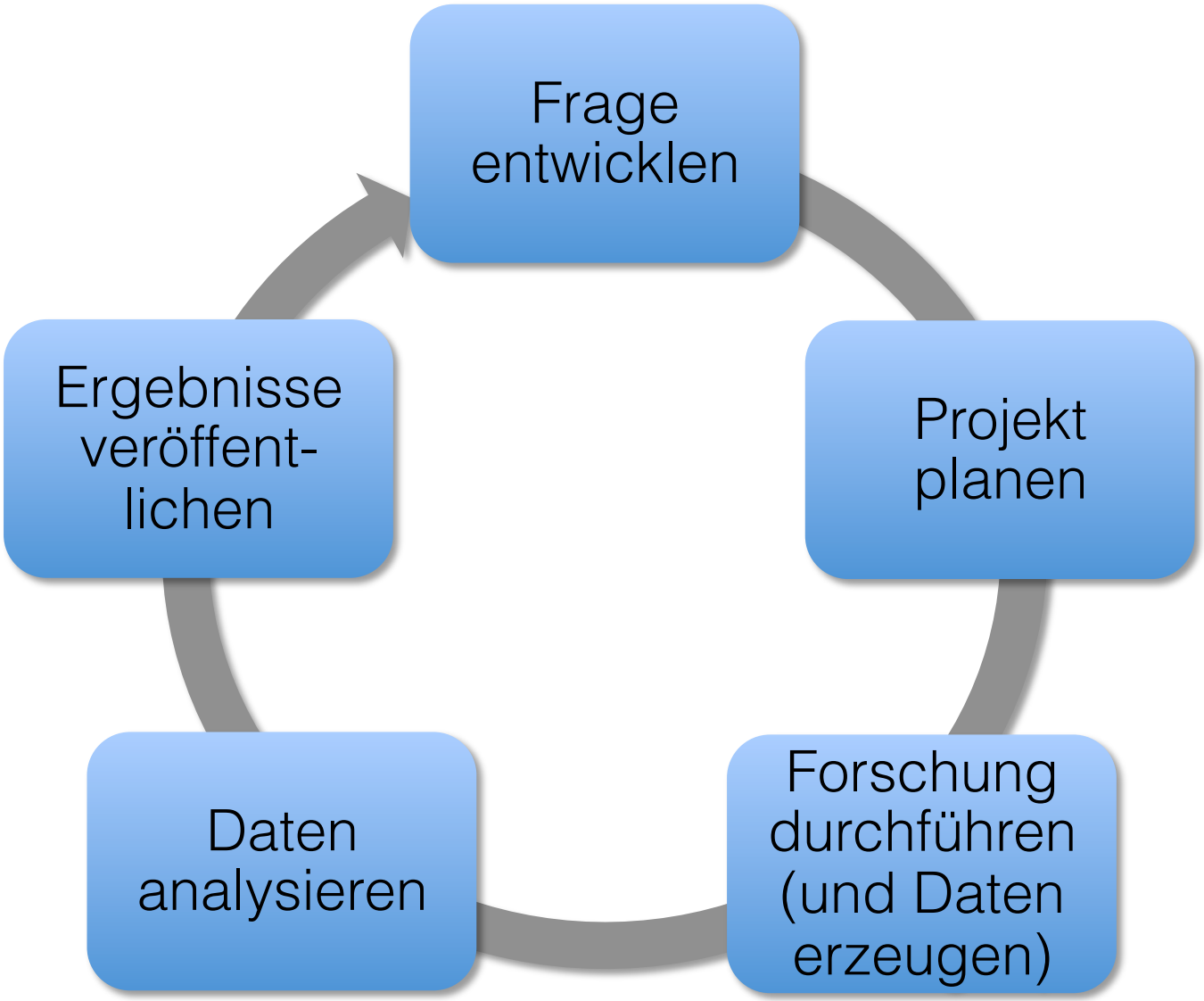
"1. Data centres are a success story for their users, and funders and policy-makers should continue to support and promote existing national data centres.

2. Data centres are important both for reference purposes, and for novel research. [...]

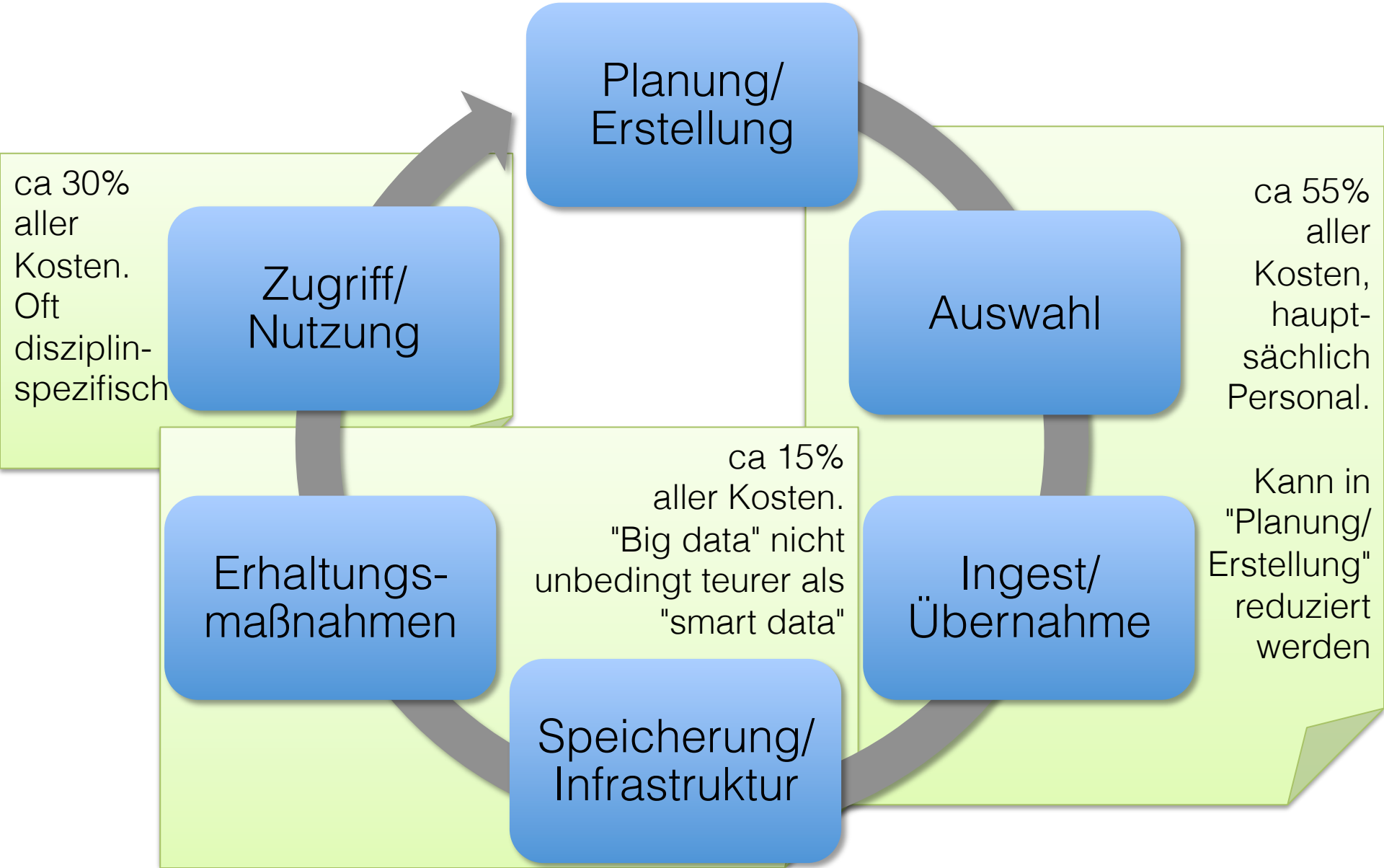
5. Although deposit levels are promising, researchers need more encouragement to deposit data. [...]

7. The national data centres are just one part of a broader landscape for data curation and storage. Further work needs to be done to investigate how they can work most effectively with local, national and international services."

Idealisierter Forschungszyklus



Infrastruktursicht: Datenzyklus



Kostenimplikationen des Datenzyklus für Service-Gruppen

Dokumentation ist (pro Datensatz) vergleichsweise günstig und zeitlich befristet, aber hohes Datenvolumen

Nachnutzung ist teuerer (aber günstig im Vergleich zur Datenerzeugung) und oftmals zeitlich unbefristet, aber geringeres Datenvolumen

Verbesserte Datennutzung verlangt relativ viel Beratung und Schulung, aber nur kurzfristig

Potential des Datenmanagements *im* Projekt

embedded data manager (z.B. INF-Projekte in SFBs)

Chance im Bereich Planung/Erstellung einzugreifen

Verlagert einige Kosten der Nachnutzung in die Phase der Datenerstellung

Vermutlich nicht insgesamt günstiger, aber bringt Zusatznutzen und erleichtert Akzeptanz

Vielen Dank!