Forschungsdaten

Einführung und ausgewählte Projekte

Christof Mainberger
Bibliotheksservice-Zentrum Baden-Württemberg

Ausgangspunkt

Bibliotheksservice-Zentrum Baden-Württemberg

- Wissenschaft basiert auf Daten

  …die objektiv und systematisch erhoben werden.

- Digitalisierung der Wissenschaft:
  - Ablösung etablierter Techniken der Datenerhebung, -dokumentation und -analyse
  - Neue „empirische“ Gebiete
  - Dynamik der Werkzeuge, Formate und Daten
  - Enormer Zuwachs der Datenmenge
  - Chance und Problem:

    Nachnutzbarkeit und Nachprüfbarkeit

nature: Availability of Data

Bibliotheksservice-Zentrum Baden-Württemberg

Positionspapiere

Bibliotheksservice-Zentrum

Diskussion

- Deutsche Forsc … G), 1998
  - „Primärdaten als … ungen sollen auf haltbaren und ge … ution, wo sie entstanden sind, … den."
  - „Auf die Aufzeichn … können, ist schon aus Gründer … Gruppe ein zwingendes Gebot. … n veröffentlichte Resu … lt werden."

http://is.gd/bGt3a

HELMHOLTZ GEMEINSCHAFT · OPEN ACCESS

Heinz Pampel | Göttingen, 27.04.2010

Diskussion

- Deutsche Forschungsgemeinschaft (DFG), 1998
  - „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."

http://is.gd/bGt3a

HELMHOLTZ | GEMEINSCHAFT OPEN ACCESS

(cc) BY

Heinz Pampel | Göttingen, 27.04.2010

Positionspapiere

BSZ Bibliotheksservice-Zentrum

Diskussion

- Berliner Erklä
  - „Establishing o dure ideally requires the ac ry individual producer of sci ltural heritage. Open access co fic research results, raw data digital representations s and scholarly multimedia mate

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities

Preface

The Internet has fundamentally changed the practical and economic realities of distributing scientific knowledge and cultural heritage. For the first time ever, the Internet now offers the chance to constitute a global and interactive representation of human knowledge, including cultural heritage and the guarantee of worldwide access.

We, the undersigned, feel obliged to address the challenges of the Internet as an emerging functional medium for distributing knowledge. Obviously, these developments will be able to significantly modify the nature of scientific publishing as well as the existing system of quality assurance.

In accordance with the spirit of the Declaration of the Budapest Open Access Initiative, the ECHO Charter and the Bethesda Statement on Open Access Publishing, we have drafted the Berlin Declaration to promote the Internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, funding agencies, libraries, archives and museums need to consider.

Goals

Our mission of disseminating knowledge is only half complete if the information is not made widely and readily available to society. New possibilities of knowledge dissemination not only through the classical form but also and increasingly through the open access paradigm via the Internet have to be supported. We define open access as a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community.

In order to realize the vision of a global and accessible representation of knowledge, the future Web has to be sustainable, interactive, and transparent. Content and software tools must be openly accessible and compatible.

Definition of an Open Access Contribution

Establishing open access as a worthwhile procedure ideally requires the active commitment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.

http://is.gd/bGtnU

(cc) BY

HELMHOLTZ | GEMEINSCHAFT OPEN ACCESS

Heinz Pampel | Göttingen, 27.04.2010

Diskussion

- Berliner Erklärung, 2003
  - „Establishing open access as a worthwhile procedure ideally requires the active commitment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.“

http://is.gd/bGtnU

HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

(cc) BY

Heinz Pampel | Göttingen, 27.04.2010

Positionspapiere

Bibliotheksservice-Zentrum

Diskussion

- OECD, 2007
  - „These Principles and commendations to the governmental science countries on access to research data from pu note data access and sharing among researc l research agencies, while at the same time, reco ious national laws, research policies and o ntries."
  - „The ultimate goal of th ove the efficiency and effectiveness of the glob d to hinder its development with onero se new costs on national science systems."

http://is.gd/bH1p7

HELMHOLTZ | GEMEINSCHAFT OPEN ACCESS

(cc) BY

Heinz Pampel | Göttingen, 27.04.2010

Diskussion

- OECD, 2007
  - „These Principles and Guidelines [...] provide broad policy recommendations to the governmental science policy and funding bodies of member countries on access to research data from public funding. They are intended to promote data access and sharing among researchers, research institutions, and national research agencies, while at the same time, recognising and taking into account, the various national laws, research policies and organisational structures of member countries."
  - „The ultimate goal of these Principles and Guidelines is to improve the efficiency and effectiveness of the global science system. They are not intended to hinder its development with onerous obligations and regulations or impose new costs on national science systems."

http://is.gd/bH1p7

HELMHOLTZ GEMEINSCHAFT OPEN ACCESS

(cc) BY

Heinz Pampel | Göttingen, 27.04.2010

Positionspapiere

Bibliotheksservice-Zentrum

Diskussion

- Schwerpunktin … tion“, 2008
  - „Es ist unbestreitbar, … tiv kurzen Phase der Auswertung durch Ei … ssen oder gar dem Verfall ausgesetzt sind. Hier … einen dringenden Handlungsbedarf hins … Archivierung und Bereitstellung dieser D … “
  - Die Aktivitäten der Alli … ichtet:
    - Primärdaten-Policy
    - Zusammenarbeit zv … Informationswissen
    - Vernetzte Repositor … gsprimärdaten

Berlin, 11.Juni 2008

Schwerpunktinitiative „Digitale Information“ der Allianz-Partnerorganisationen

Alexander von Humboldt-Stiftung
Deutscher Akademischer Austauschdienst
Deutsche Forschungsgemeinschaft
Fraunhofer-Gesellschaft
Helmholtz-Gemeinschaft Deutscher Forschungszentren
Hochschulrektorenkonferenz
Leibniz-Gemeinschaft
Max-Planck-Gesellschaft
Wissenschaftsrat

http://is.gd/bH1vF

HELMHOLTZ | GEMEINSCHAFT ····· OPEN ACCESS

(cc) BY

Heinz Pampel | Göttingen, 27.04.2010

Diskussion

- Schwerpunktinitiative „Digitale Information“, 2008
  - „Es ist unbestreitbar, dass viele dieser Daten nach einer relativ kurzen Phase der Auswertung durch Einzelne oder kleine Gruppen dem Vergessen oder gar dem Verfall ausgesetzt sind. Hier sehen alle Wissenschaftseinrichtungen einen dringenden Handlungsbedarf hinsichtlich der systematischen Sicherung, Archivierung und Bereitstellung dieser Daten für die Nachnutzung durch Dritte.“
  - Die Aktivitäten der Allianz-Initiative sind auf drei Bereiche gerichtet:
    - Primärdaten-Policy
    - Zusammenarbeit zwischen Fachwissenschaftlern und Informationswissenschaftlern anzustoßen
    - Vernetzte Repositorien- und Archivstrukturen für Forschungsprimärdaten

http://is.gd/bH1vF

HELMHOLTZ | GEMEINSCHAFT OPEN ACCESS

(cc) BY

Heinz Pampel | Göttingen, 27.04.2010

Diskussion

Bibliotheksservice-Zentrum Baden-Württemberg

… und vieles, vieles anderes:

- 2009: Unterausschuss für Informationsmanagement der DFG
  DFG Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung von Forschungsdaten
- 2009: Kapitel im nestor Handbuch
- 2010: GWK Kommission Zukunft der Infrastruktur („KII“)
  AG zu Forschungsdaten: Empfehlungen an die GWK
- 2010: Ausschreibung der DFG:

  Informationsinfrastrukturen für Forschungsdaten
- 2010: Ergänzung des Antragsformular der DFG:
  … um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen

und vieles disziplinspezifisch und international…

- z.B. vorgestern: zwei Veröffentlichungen des ODE-Project

DINI-Positionspapier

Bibliotheksservice-Zentrum Baden-Württemberg

- **Disziplinarität**
  Forschungsdaten variieren nach Disziplin
- **Organisation**
  Rahmen für Forschungsdatenmanagement: Bibliotheken, Daten- und Rechenzentren unterstützen Forschende
- **Zugang**
  Via Nachweisdienste, Kataloge.
  Open Access, soweit rechtlich möglich
- **Qualität**
  Standards, Formale Erschließung, persistente Adressen
- **Technik**
  Vertrauenswürdige Archive, interoperable Schnittstellen

Erschließung und Nachweis

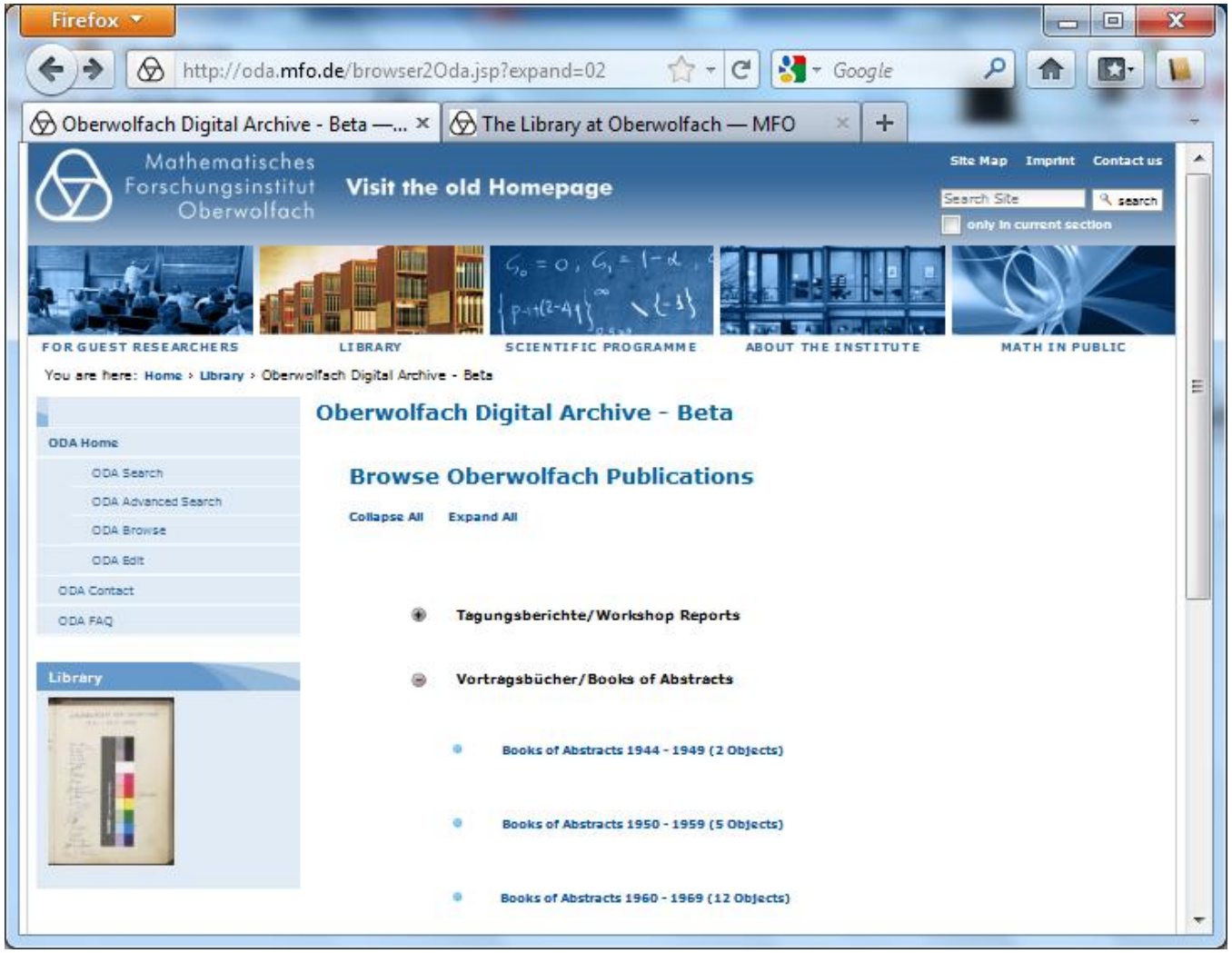
Bibliotheksservice-Zentrum Baden-Württemberg

- Formate:
  - DFG 2009: mindestens DC (simple)
  - OECD: Marc-Records
  - Museen: CIDOC (CRM) als Rahmen
  - Technische / administrative Metadaten
- Persistente Adressierung:
  - DOIs, URNs, aber in welcher Granularität?
- Recherche:
  - OPACs, SWB, spezielle Dienste (z.B. DataCite)?

Oberwolfach Digital Archive

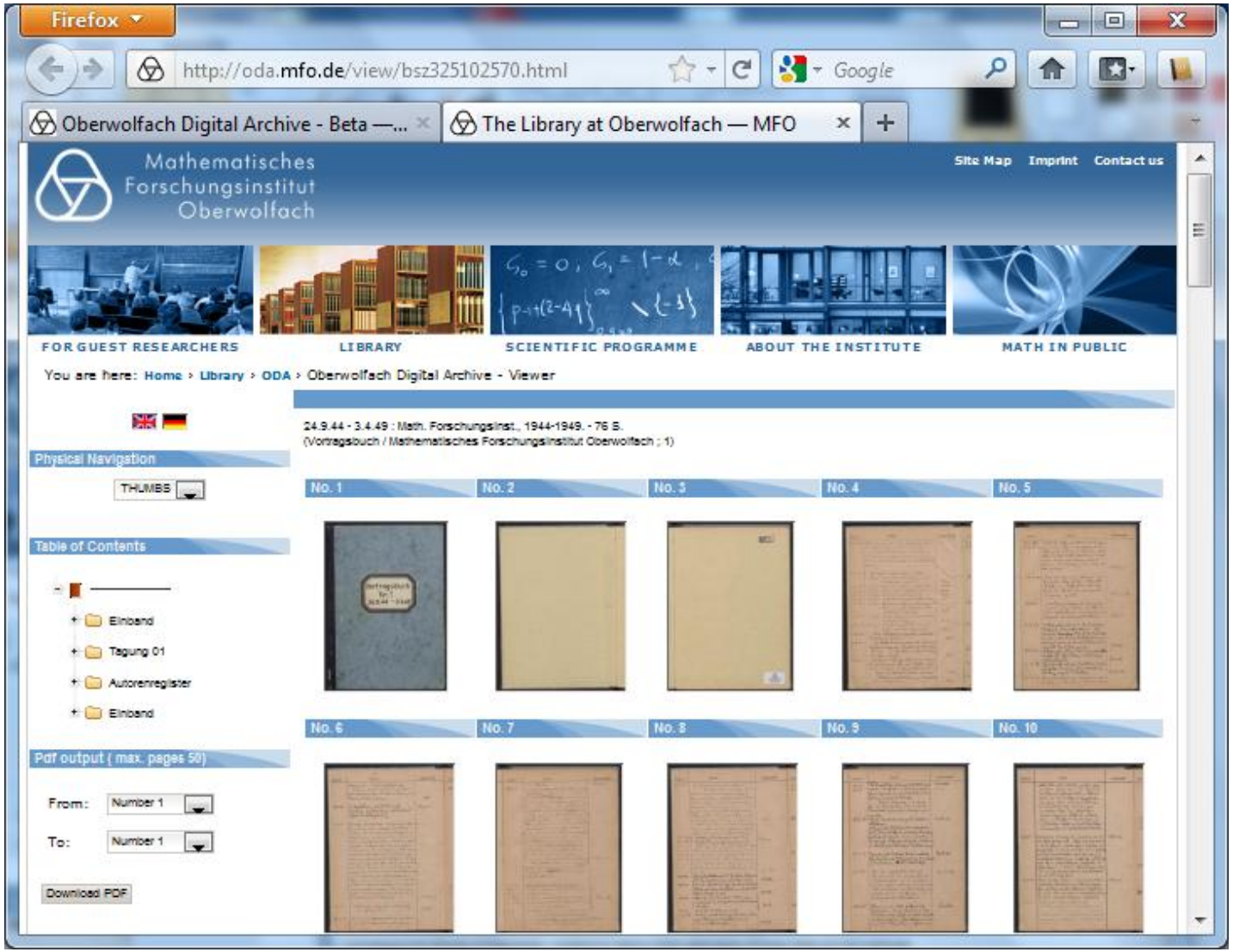
Bibliotheksservice-Zentrum Baden-Württemberg

Oberwolfach Digital Archive

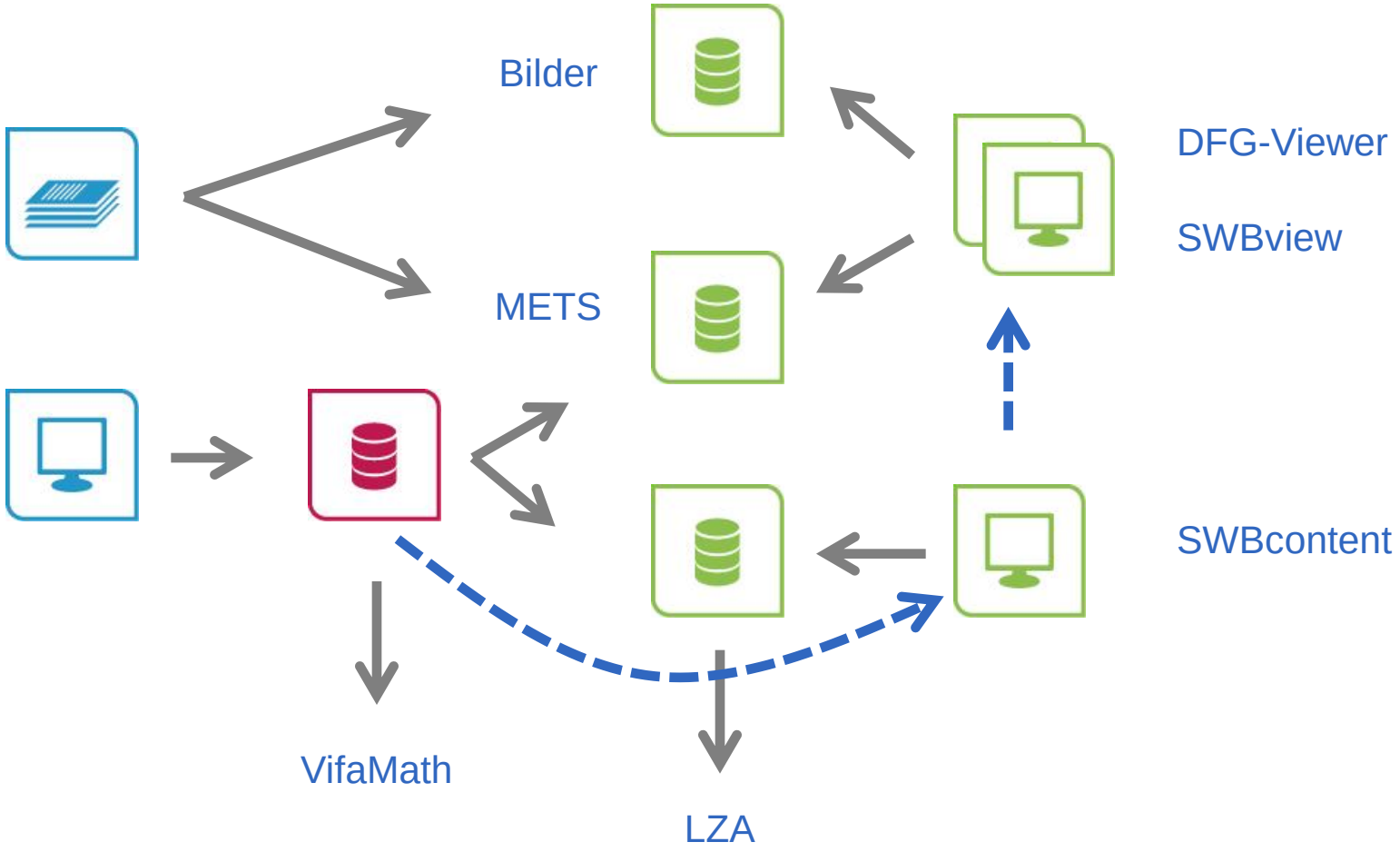


Oberwolfach Digital Archive

Bibliotheksservice-Zentrum Baden-Württemberg



Oberwolfach Digital Archive



Was soll archiviert werden?

Bibliotheksservice-Zentrum Baden-Württemberg

- Forschungsdaten sind dynamisch:
  - Sie werden in vielen Schritten bearbeitet
  - Sie werden aus diversen Quellen aggregiert
  - Sie sind (oft) nicht standardisiert
- Nachnutzung bedingt Standards und Offenheit
  - Z.B. TIFF, XML, Open Linked Data, …

- Nachnutzung vs. Nachprüfung:

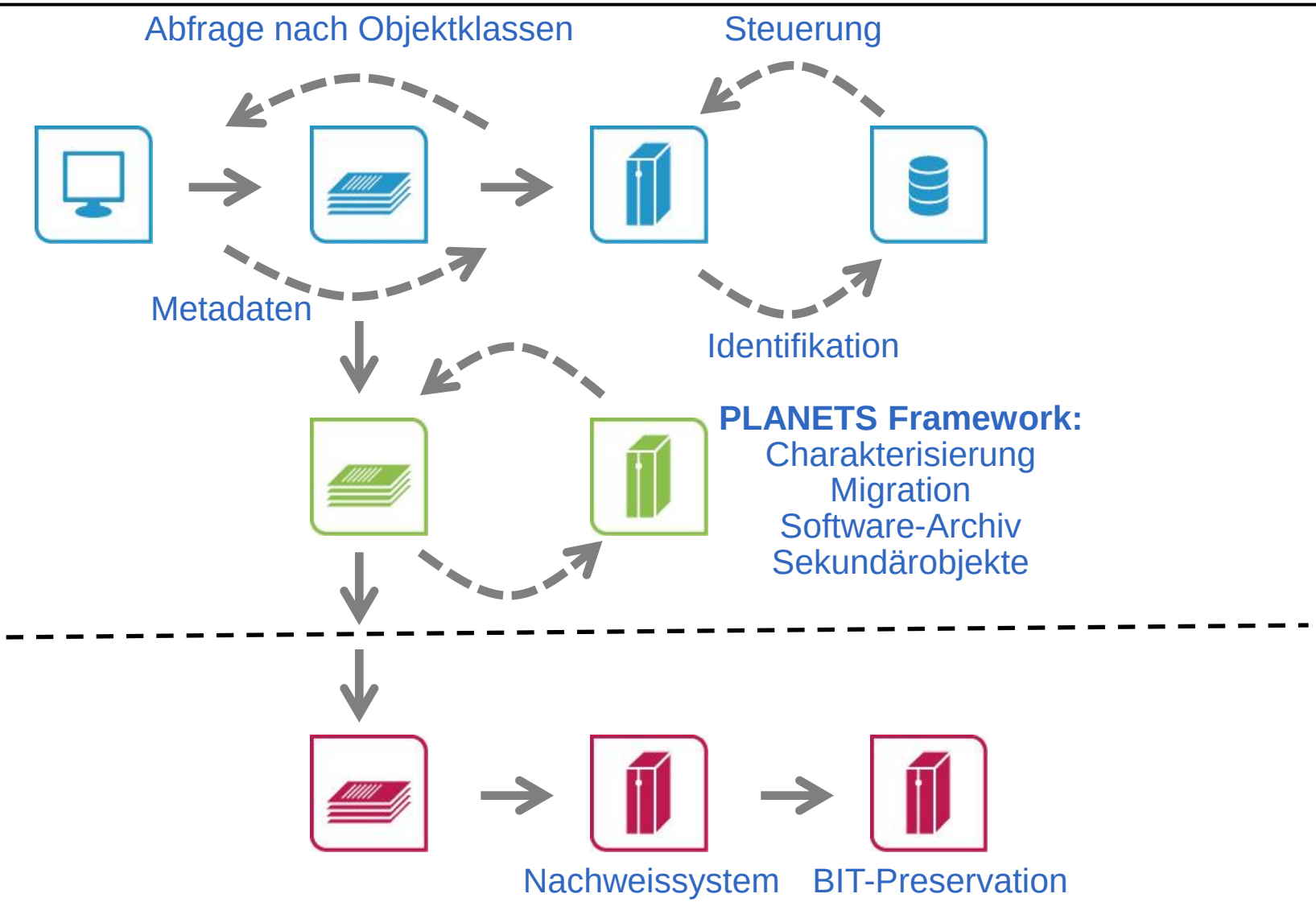
  Was soll archiviert werden??

bw-fla: LZA von Daten

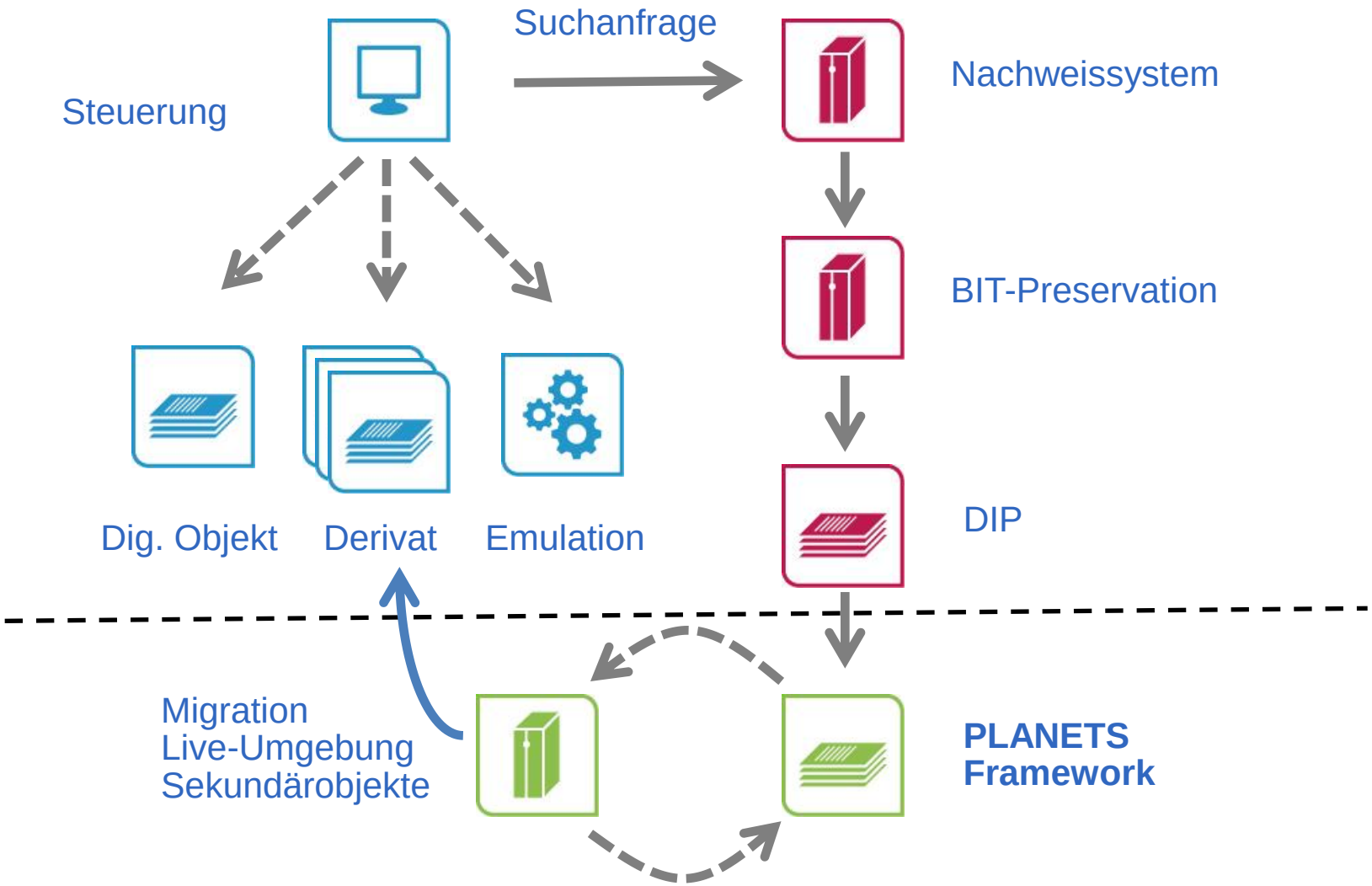
Bibliotheksservice-Zentrum Baden-Württemberg

- Komplexe dynamische Objekte
  - Wissenschaftliche Primärdaten
  - Komplexe Arbeitsumgebungen
  - Datenbanken
  - Verwaltungsvorgänge
- Unterstützung beim Ingest und Access
- Sowohl Migration, als auch Emulation
- Metadaten zur technischen Ablaufumgebung
- Einbettung in bestehende Strukturen
- RZ Freiburg, KIZ Ulm, RZ Mannheim, RUS, BSZ, LA

bw-fla Ingest



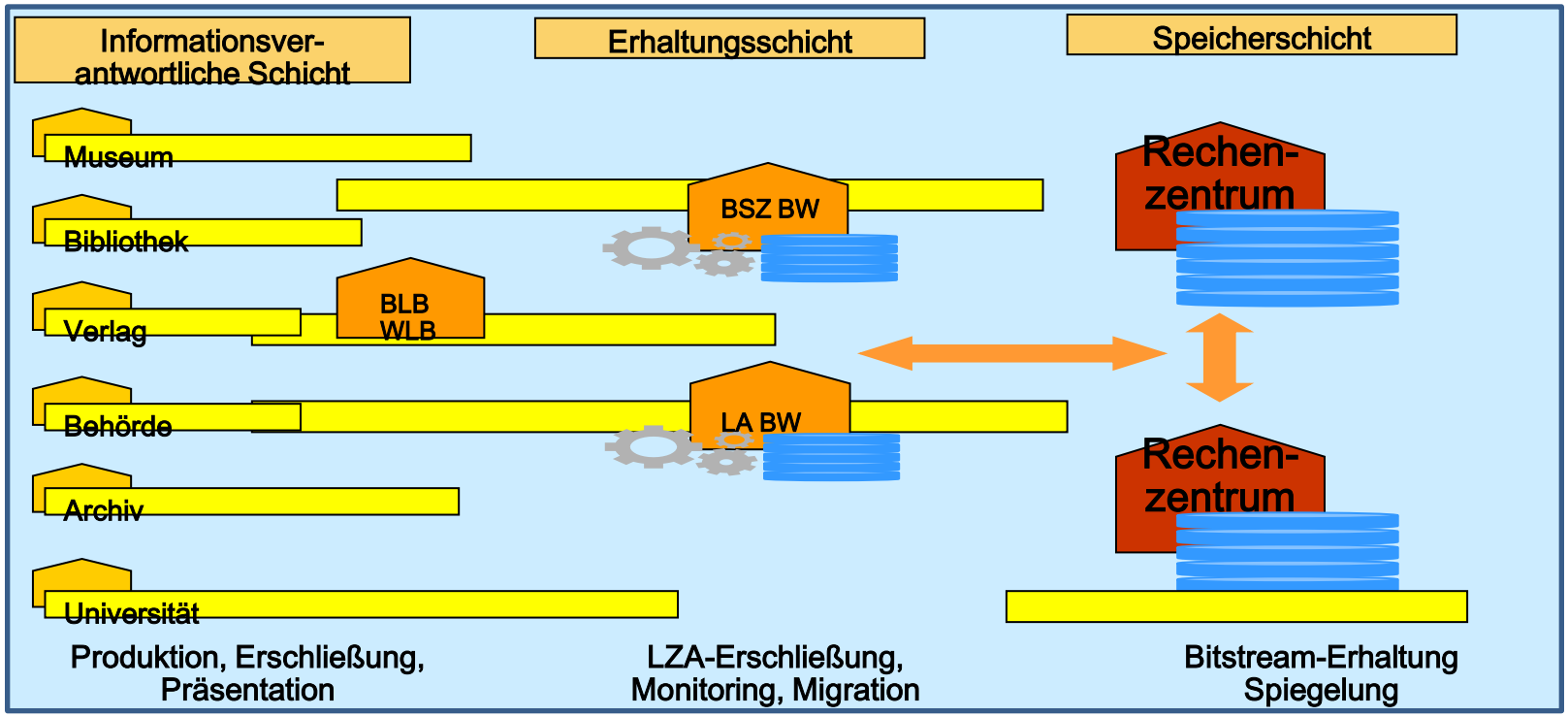
bw-fla Access



Festlegung der Organisation

Bibliotheksservice-Zentrum Baden-Württemberg

- Analog des Schichtenmodells zur digitalen LZA?



- Oder disziplinär, z.B. in Weltdatenzentren?
- Oder beides?

Rechte und Anreize

Bibliotheksservice-Zentrum Baden-Württemberg

- Freiheit der Wissenschaft (Art. 5 GG):
  Freiheit, Forschungsdaten nicht freizugeben?
  Wem gehören eigentlich die Daten?
- Berücksichtigung einer exklusiven Nutzung
  Durch Moving Walls, Zugriffsbeschränkung, Lizenzen, etc.
- Forderung der Förderer:
  nach einem Forschungsdatenmanagement
  – und dessen Finanzierung…
- Kriterium für leistungsbezogene Mittelvergabe
  für qualitätsgesicherte, publizierte Forschungsdaten
- Erarbeitung einer Forschungsdatenpolicy
  z.B. University of Edinburgh
- Integration in Forschungsinformationssystme

Aufgabe von Bibliotheken

Bibliotheksservice-Zentrum Baden-Württemberg

- Bereitstellung qualifizierten Personals

- Erarbeitung von Workflows und Best Practices
- Festlegung von Metadaten
- Beschreibung des rechtlichen Rahmens
- Bereitstellung von Nachweisinstrumenten
- Zugang zu technischer Infrastruktur

und außerdem: Überreden und Überzeugen…

PURDUE UNIVERSITY **Libraries** *Access. Knowledge. Success.*

INTRODUCTION

Librarians at Purdue University are beginning to identify the scientific datasets that are being generated by our faculty and researchers as information assets to be collected, preserved, and made accessible as a function of the library's collection development. These librarians are subject-area specialists, and many have advanced degrees in their respective disciplines in addition to a degree in library science. They have all been trained in collection management; however, much of this training was related to traditional formats such as monographs and serials and not datasets. In our experience, one of the most effective tactics for eliciting datasets for the collection is a simple librarian-researcher interview. In this poster, we share a set of ten questions that a librarian can use as a starting point for such a "data interview". It is not a comprehensive strategy but instead a practical tool to draw out information that needs to be considered in order to evaluate the suitability of a dataset for the collection and the requirements for the infrastructure and services that will be needed for data curation.

#1 What is the story of the data?

Begin the interview with an open-ended question that allows the researcher to talk freely about his or her research, scientific workflow, and community of practice. This leads some insight into the value of the dataset and how it may fit into your collection and be used, and it provides the *context* for understanding how and why the dataset was created and how it was processed and analyzed.

#2 What form and format are the data in?

What computing environments (e.g., software) are required to use the data? If the data are in proprietary structures, you may consider reformatting them into agnostic formats or ones that can be more *easily re-versioned*. Is there any existing *metadata*, either external to the data or description that could be extracted from it? Ideally the data could be described to be discoverable by researchers from another discipline.

#3 What is the expected lifespan of the dataset?

In many cases, there are distinctions in the utility of a dataset as it begins in a raw state and then is analyzed and processed into new forms and versions as a result of different steps in the research workflow. Different entities may have custody of the data and use it for different purposes at different times, affecting its *provenance*. Funding agencies may require that data be archived for a prescribed period of time or you may forecast its future value and the amount of time it should be retained. The data may be described and archived for effective *preservation* to ensure its accessibility and integrity over time.

#4 How could the data be used, reused, and repurposed?

This is a primary *selection* criterion that also impacts how the data are *accessed* and what *policies* may be needed to govern its use. As data are archived and shared, new and unintended uses for the data may increase its value. For example, a research dataset may be repurposed as a learning object.

#5 How large is the dataset, and what is its rate of growth?

It is important to quantify the size of the data for storage and network provisioning if you intend to *ingest* it into your repository. What is its physical (bits) and logical (records) *scale*? Is the dataset static or dynamic? Ask for a sample of the data to examine.

#6 Who are the potential audiences for the data?

Information regarding potential users of the data and the users' needs is paramount. Along with potential uses for the data, this is another primary *selection* criterion. In some cases, the data may need to be embargoed or restricted to a limited group of users who are granted *permission* to access it.

#7 Who owns the data?

Establishing and maintaining the *intellectual property* represented by the data should be discussed at the earliest opportunity, and any conflicts should be resolved up-front. Many organizations have a submission policy that asks the contributor to verify that they own the data and have the right to submit it.

#8 Does the dataset include any sensitive information?

All data should be reviewed for information that violates *confidentiality*, such as identification information on human subjects. Data curation activities should be informed by institutional review board requirements.

#9 What publications or discoveries have resulted from the data?

The researchers may have a bias regarding the importance of their data. The purpose of this question is to establish an objective metric for determining the value of the data for the collection. Different metrics may be more appropriate in determining the *selection* criteria for different kinds of data and data collections.

#10 How should the data be made accessible?

There is value in making data accessible using a conventional web-based user interface, but machine-to-machine interfaces should also be evaluated. These *methods of access* will be informed by the answers to the previous questions, and this question can be asked in an open-ended manner to fill in any gaps remaining at the conclusion of the interview.

SUMMARY

Although building robust collections of datasets present several complexities and challenges to resolve, the process of looking at scientific datasets as information assets and exploring what is needed to develop and manage data collections is similar to the traditional collection development practices that have been successfully employed by librarians for decades. We offer these ten "data interview" questions as a springboard for librarians to explore data curation in greater depth and specialization.

Michael Witt (mwitt@purdue.edu)
Assistant Professor of Library Science

Jake Carlson (jrcarlso@purdue.edu)
Data Research Scientist

Purdue University Libraries
Distributed Data Curation Center
http://d2c2.lib.purdue.edu

D2C2 Distributed Data Curation Center

"Conducting a Data Interview"

Michael Witt & Jake Carlson, Purdue University Libraries, West Lafayette, Indiana, USA