Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

Metadaten für geisteswissenschaftliche Forschungsplattformen: Welche Standards mit welchem Nutzen?

SAGW-Tagung

Prof. Dr. Lukas Rosenthaler
Data & Service Center of the Humanities
Digital Humanities Lab / Universität Basel

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

Warum?

Was ist wichtig für geisteswissenschaftliche Forschungsplattformen?

- “open data”: Daten sollen “offen” sein

  überprüfen der Resultate der Forschung

  neue Erkenntnisse aus bestehenden Daten gewinnen 👉 (*re-use*)

- FAIR:

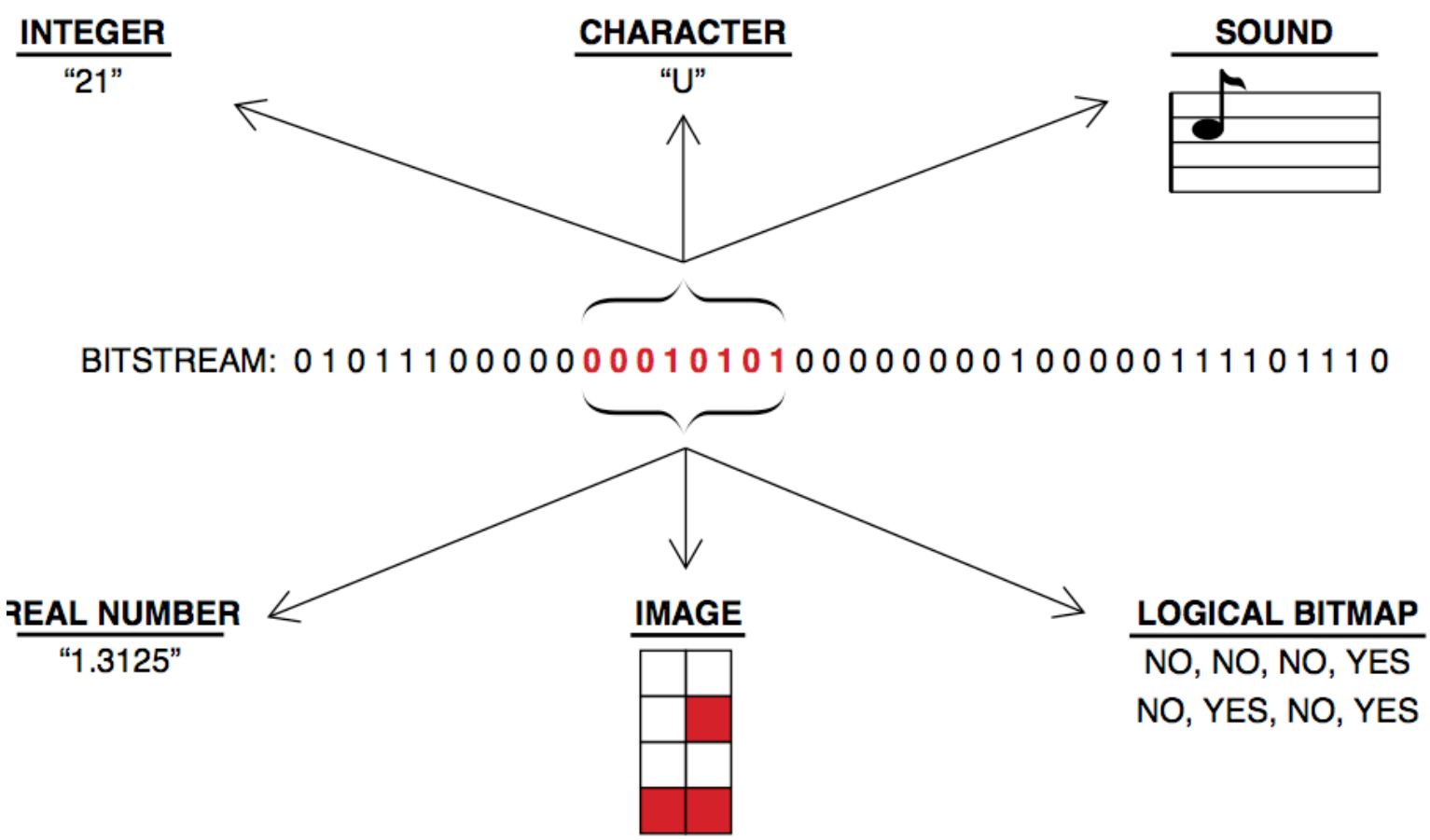
  Findable, ***A***ccessible, ***I***nteroperable, ***R***eusable

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

Digitale Daten

Bedeutung der Bits nicht automatisch gegeben:
Das ***Format*** definiert die Bedeutung!



Data & Service Center for the Humanities DaSCH

Beispiel: TIFF

Digital Humanities Lab

Section 2: TIFF Structure

TIFF is an image file format. In this document, a *file* is defined to be a sequence of 8-bit bytes, where the bytes are numbered from 0 to N. The largest possible TIFF file is 2**32 bytes in length.

A TIFF file begins with an 8-byte *image file header* that points to an *image file directory* (*IFD*). An image file directory contains information about the image, as well as pointers to the actual image data.

The following paragraphs describe the image file header and IFD in more detail.

See Figure 1.

Image File Header

A TIFF file begins with an 8-byte image file header, containing the following information:

Bytes 0-1: The byte order used within the file. Legal values are:

"II" (4949.H)

"MM" (4D4D.H)

In the "II" format, byte order is always from the least significant byte to the most significant byte, for both 16-bit and 32-bit integers This is called *little-endian* byte order. In the "MM" format, byte order is always from most significant to least significant, for both 16-bit and 32-bit integers. This is called *big-endian* byte order.

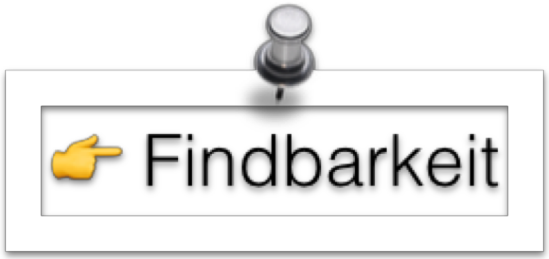
Bytes 2-3 An arbitrary but carefully chosen number (42) that further identifies the file as a TIFF file.

The byte order depends on the value of Bytes 0-1.

Bytes 4-7 The offset (in bytes) of the first IFD. The directory may be at any location in the file after the header but *must begin on a word boundary*. In particular, an Image File Directory may follow the image data it describes. Readers must follow the pointers wherever they may lead.

Metadaten: Daten über Daten

- ***technische*** Metadaten
  - Formate
  - Angaben zur Digitalisierung (Farbprofile etc.)
  - …
- ***beschreibende*** Metadaten:
  - *Sekundärdaten*, welche wichtig zur richtigen Interpretation der *Primärdaten* sind
    - Herkunft (Provenienz)
    - Taxanomien
    - Rechte
    - …



Auch Metadaten sind digitale Daten (mit allen Problemen)!

Zugriff auf die Daten

- für die/den ForscherIn
  👉 einfache Benutzeroberfläche

- maschinenlesbar
  👉 API (***A***pplication ***P***rogrammer ***I***nterface)

- Sowohl für die

  primären Daten

  Metadaten (sekundäre Daten)

Interoperabilität = Standard

- ***Interoperabilität*** bedingt, dass sich Daten ***Anbieter*** und Daten ***Nutzer*** auf drei Ebenen einigen:

  Bedeutung der Bits:
  Formate 👉 **Syntax**

  Interpretation der Metadaten:
  Schemas 👉 **Semantik**

  Zugriffsart 👉 **API**

- ***Standards***: weitgehende akzeptierte Übereinkunft, wie "etwas" implementiert wird…

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

Anforderungen an Standards in den GW

- grosse ***Akzeptanz***
- ***offene*** Definition
- ***Einfachheit*** (KISS)

  keep it simple, stupid!
- Potential für ***Langlebigkeit***
  (“Halbwertszeit” der Daten wird in Dekaden oder Jahrhunderten gemessen)

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

1. Ebene: Primärdaten

- Abhängig vom Medientyp:
  - ➡️ Text: XML/TEI, UTF-8/16/32, LaTeX,
  - ➡️ Bild: JPEG2000, TIFF (⚠️), ~~JPEG~~
  - ➡️ Ton: WAV, FLAC, Ogg
  - ➡️ Film, Video: ⚠️ DCP?, MP4?
  - ➡️ Datenbanken: ~~FileMaker~~, ~~MSAccess~~, ~~XXSQL~~
    👉 RDF (Resource Description Framework), a W3C standard

2. Ebene: Sekundärdaten

- Weitverbreitete Standards:

  Dublin Core, das KGV der Metadaten
  (KGV: *K*leinster *G*emeinsamer *N*enner)

  METS: Kodierung ***deskriptiver***, ***administrativer*** und ***struktureller*** Metadaten für Objekte in einer ***digitalen Bibliothek***

  CIDOC: ***Conceptual Reference Model*** (CRM): Definitionen und formale Struktur zur Beschreibung von ***impliziten*** and ***expliziten Konzepten*** und ***Beziehungen*** im Bereich der ***Dokumentation von kulturellem Erbe***

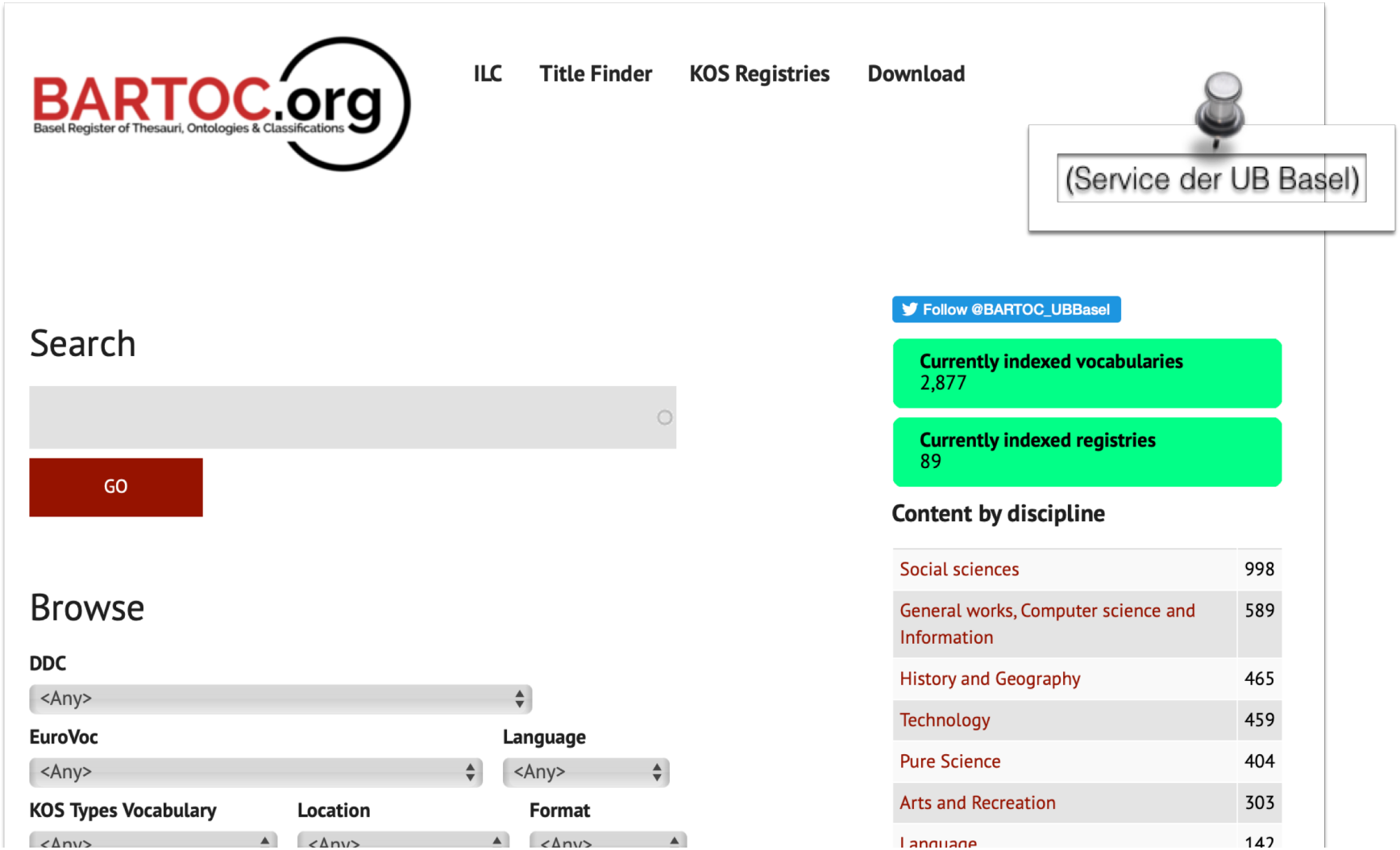
  …und viele mehr, abhängig von der Disziplin…

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

Übersicht über Metadatenschemen und Thesauri

http://bartoc.org



Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

3. Ebene: Zugang (API)
(Datenobjekte)

- ***International Image Interoperability Framework (IIIF)***

  weit akzeptierter Standard für den Zugriff auf ***bildhafte*** Objekte & **Interoperabilität**

  Erweiterung für ***andere Medien*** in Arbeit

  Beruht auf *HTTP*-Protokoll (Webtechnologie)

  “Metadaten”: *JSON-LD*

3. Ebene: Zugang (API)

(Daten, Metadaten,…)

- *Protokoll*: **REST** (***Re***presentational ***S***tate ***T***ransfer):
  - beruht auf Web-Standard 👉 HTTP-Protokoll:
    GET, PUT, POST, DELETE,…
- *Datentransfer-Format*:
  - **JSON** (***J***ava***S***cript ***O***bject ***N***otation):
    weit akzeptierte, einfache, effiziente Datenbeschreibung
  - **JSON-LD**: Variante für ***Linked Open Data***
  - **XML**: e***X***tensible ***M***arkup ***L***anguage

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

4. Ebene: Verknüpfungen

(👉Datenbanken)

- Datenbanken verknüpfen & aggregieren

  primäre Daten (Datenobjekte)

  sekundäre Daten (alle Varianten von Metadaten)

- sind ***komplexe “Objekte”***

- sollten auch ***API*** haben

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

Semantic Web

RDF, RDFS, OWL, JSON-LD = Linked Open Data (LOD)

- verbindet die **Repräsentation** von *Daten*, *Metadat* **...nter einem Dach**

- Unterstü ...dards
  👉 "Onto

- Standardisierte **Repräsentation** (...Formate) und ***Zugriff*** (*API* und *Datentransfer-Format*)

- *offener, gut dokumentierter Standard* (W3C)

👉grösstmögliche **Offenheit** und **Interoperabilität**

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

Swiss Art Research Infrastructure (SARI)

- **SARI** ist eine nationale Forschungsinfrastruktur, welche einen ***einheitlichen*** und ***gegenseitigen*** Zugang zu Forschungsdaten von spezialisierten Institutionen im Bereich Kunstgeschichte ermöglicht.§
- **SARI**s Netzwerk ist forschungsbezogen und baut für den Zugriff auf die Daten und Datenobjekte auf offenen Standards (***RDF***, ***RDFS***, **OWL**, ***IIIF***, …) 👉 ***LOD*** + ***IIIF***
- um diese Ziele zu erreichen beteiligt sich SARI an der *Entwicklung* und *Erweiterung* von *domänen-spezifischen Metadatenstandards* (Ontologien), z.B. ***CIDOC-CRM***, um das Potential von LOD voll auszuschöpfen (“linked data” statt Datensilo)
- Um den ***langfristigen Zugriff*** auf die Forschungsdaten gemäss den FAIR-Prinzipen zu garantieren, arbeitet SARI mit dem DaSCH zusammen.

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

Data and Service Center for the Humanities

(DaSCH, eine Unternehmung der SAGW)

- Das DaSCH ist eine nationale Forschungsinfrastruktur welche den langfristigen Zugang zu komplexen, verknüpften Forschungsdaten (👉Datenbanken, qualitative Daten) und den dazugehörigen digitalen Objekten garantiert
- Das DaSCH unterstützt offene Standards und ist interoperabel auf allen Ebenen:
  - ➡️ **IIIF für Datenobjekte**
  - ➡️ **RDF, RDFS, OWL** für Daten und Metadaten
  - ➡️ Unterstützung von **Standard-Ontologien**
  - ➡️ **REST API** mit
  - ➡️ **FAIR** data

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

DaSCH Langzeitarchivierung

- periodische Speicherung auf haltbarstem Medium (Magnetband) mit periodischer Migration
- sorgfältige Wahl der Formate für digitale Objekte
- Datenstruktur und Daten in einfachstem Schema (UTF-8-Datei in RDF/N3)

  für den Menschen lesbar

  maschinenlesbar

  Datenstruktur, Daten und Metadaten im gleichen Format (RDF)

Data & Service Center for the Humanities DaSCH

Digital Humanities Lab

Fazit

- Implementierung offener, “guter” Standards ist essentiell
- **LOD** & **IIIF** bilden *unschlagbares “Team”*

  universal einsetzbar

  flexibel, offen und gut dokumentiert

  unterstützt viele Standards in *einheitlicher Umgebung*
- *Langzeit-Verfügbarkeit* stellt besondere Ansprüche!

Data & Service
Center for
the Humanities
DaSCH

Information portal:

http://dasch.swiss

Email:

info@dasch.swiss

vera.chiquet@unibas.ch