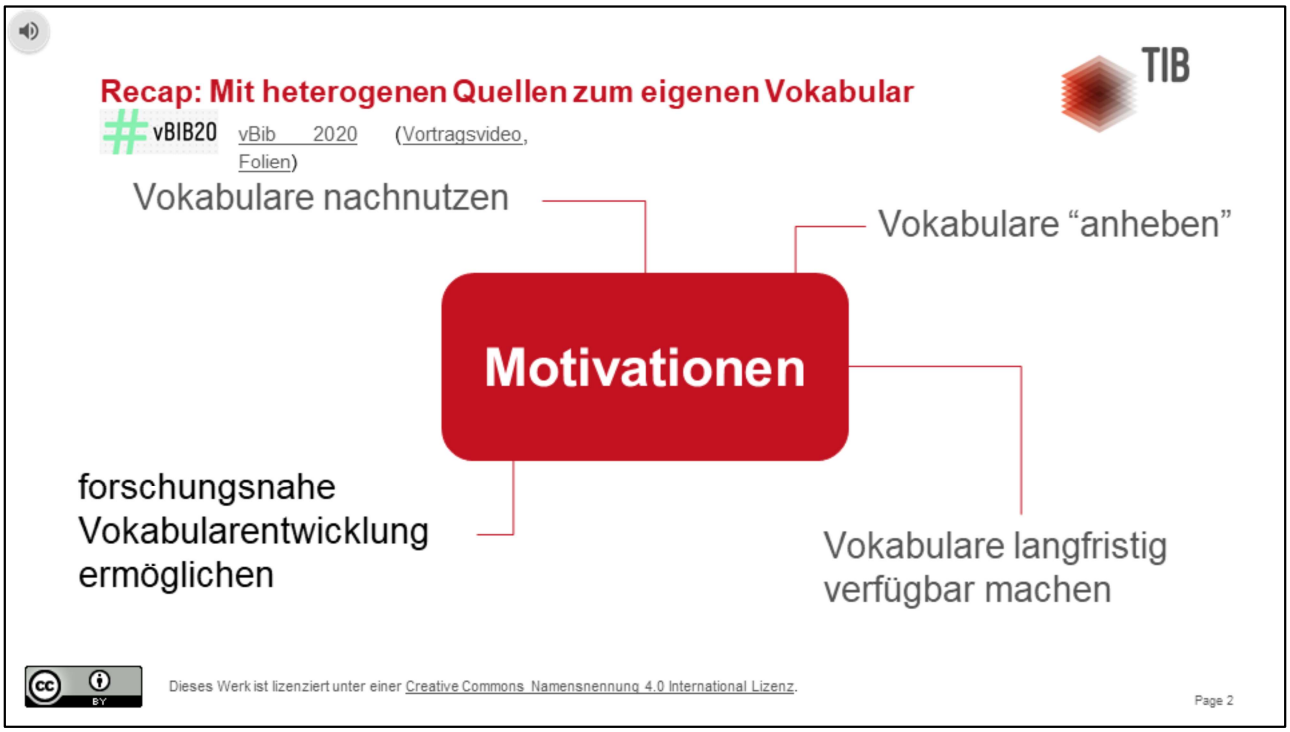
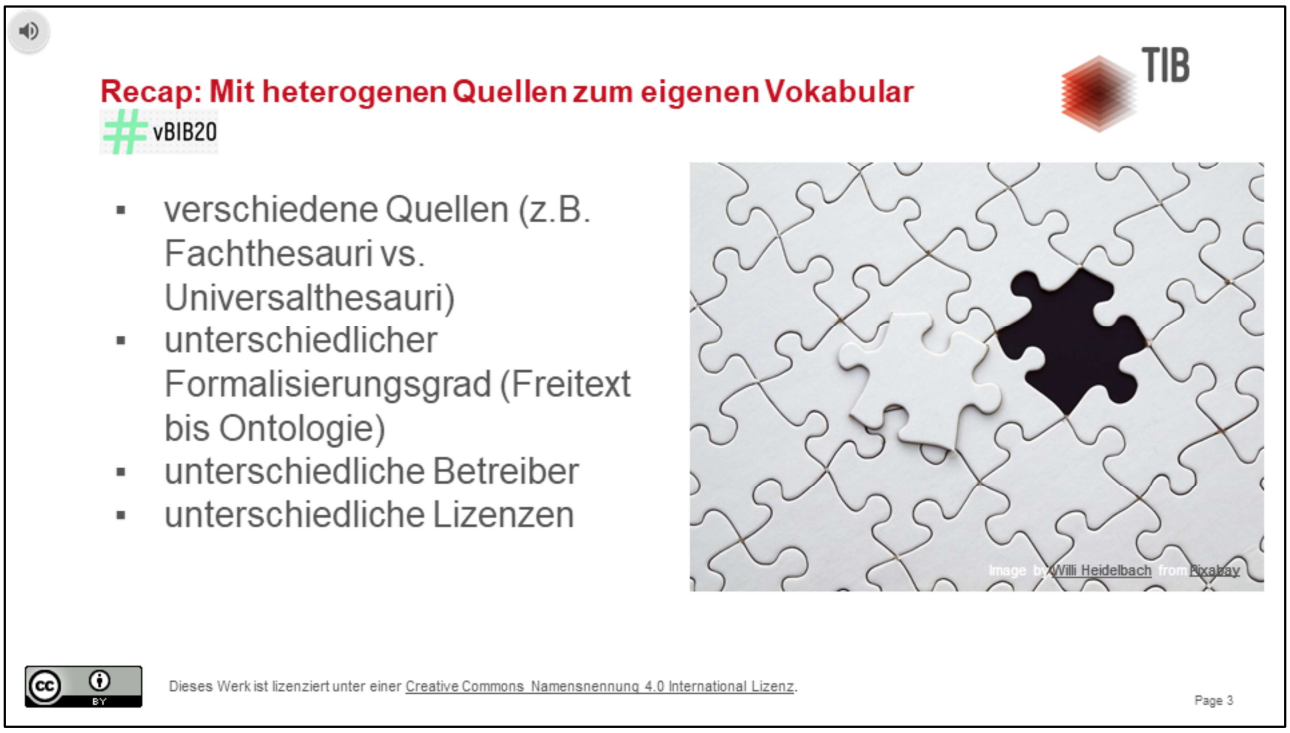

**Vortragsnotizen:**

- Willkommen :-)
- Werkstattbericht aus den Fachinformationsdiensten Mobilitäts- und Verkehrsforschung und BAUdigital (Bauingenieurwesen, Architektur und Urbanistik)
- Titel “Mit heterogenen Quellen…” > spielt auf Vortrag bei der vBIB 2020 an
- der Titel enthält bewusst nicht die Wörter Sacherschließung und GND-Arbeit, weil beide Projekte Quellen neben der GND berücksichtigen wollten
- Interesse der Thesaurus-Arbeitspakete der beiden FIDs an der GND: insbesondere Sachschlagwörter und Entitätstypen wie Verkehrsmittel, Bauwerke und Denkmäler



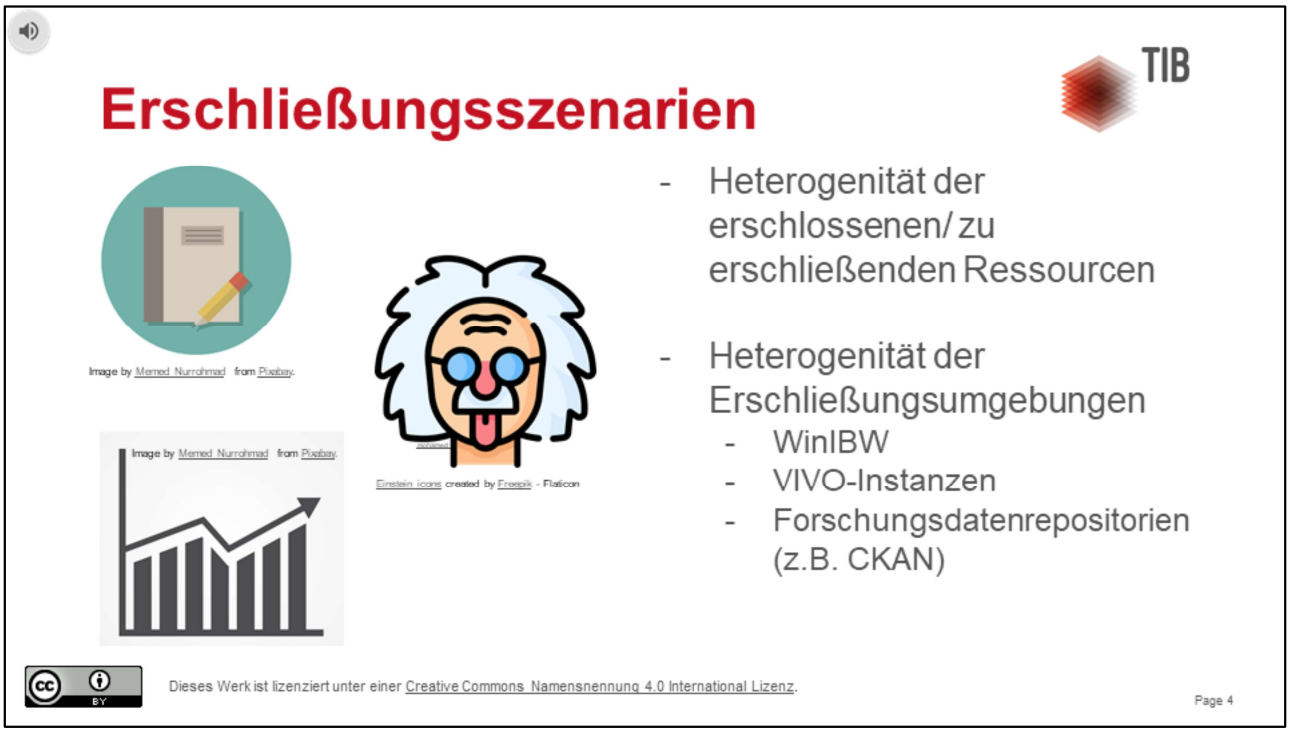

- auf der vBIB 2020 hatten wir aus Perspektive des FID Move berichtet, dass wir eine möglichst hohe Nachnutzung existierender Quellen mit fachlicher Terminologie erreichen wollten
- dieser modulare Ansatz erschien als geeignete Lösung, um einen umfassenden Fachwortschatz aufzubauen: so konnten wir unser wichtigstes Erschließungsinstrument, die GND, schnell mit anderen Vokabularen kombinieren und mit fachlichen Konzepten und Fachausdrücken anreichern, die Domänenexperten in der GND vermisst hatten
- im FID BAUdigital haben wir uns diese Nachnutzung sogar als zwingende Notwendigkeit selbst auferlegt: hier bringt nämlich einer der Partner, das Fraunhofer Informationszentrum Raum und Bau (oder kurz: IRB) verschiedene Spezialvokabulare mit, und zwar die Thesauri FINDEX Raum und FINDEX Bau sowie eine organisch gewachsene Schlagwortliste
- diese IRB-Vokabulare wollen wir langfristig versioniert verfügbar machen, sodass sie auch die FAIR-Prinzipien erfüllen
- zu guter letzt wollen wir auch eine Weiterentwicklung dieser Vokabulare aus der Fachcommunity heraus ermöglichen - sei es in Form einzelner Begriffsvorschläge oder ganzer Vokabulare
    - FIDs entwickeln Services für Forschende, deren Funktionalität teilweise sehr stark auf terminologischen Begriffssystemen beruhen >

deswegen wäre es gut, wenn Forschende ihre Terminologiebedarfe/-wünsche/ -expertise mit einbringen könnten - wie gut sich das umsetzen ließe, können wir zum jetzigen Zeitpunkt noch nicht beurteilen




- der von uns gewählte Nachnutzungsansatz hat uns mit Heterogenität in einer Reihe von Dimensionen konfrontiert:
  - potentielle Quellen haben unterschiedliche Herkunft und wurden mit unterschiedlichen Zwecken erstellt
    - Allgemeinthesauri, die viele Disziplinen abdecken, sind für interdisziplinäre Forschungsommunities eigentlich genial - lassen aber vielleicht fachliche Tiefe und Spezialterminologie vermissen
    - Fachthesauri gehen für spezialisierte Erschließungskontexte in die richtige Tiefe - sind aber oft genug noch hinter einer relativ restriktiven Lizenz versteckt und eignen sich nicht für eine freie Nachnutzung oder gar Adaptionen
  - auch die Formalisierungsgrade der verschiedenen Terminologiequellen können sich unterscheiden
    - Thesauri richten sich oft schon nach dem SKOS-Standard, nehmen also einen mittleren Formalisierungsgrad ein und sind mit Semantic-Web-Anwendungen grundsätzlich kompatibel
    - Vokabulare aus der Forschung selbst variieren im Formalisierungsgrad dagegen: von Freitext bis zur Ontologie

kann jedes Format dabei sein (lexikonähniche Artikel, Wörterbücher, Tabellen, Glossar, Word-Dokumente, PDFs, HTML-Dokumente, Wikis, etc. … )
        - die Integration dieser heterogenen Quellen erfordert teilweise Transformationen - zwar ist es immer relativ einfach jedes Vokabular nach RDF zu transformieren, aber insgesamt wird jedes Vokabular dadurch zu einem Einzelfall, den man sich genau anschauen muss
    - auch restriktive Lizenzen erschweren die Arbeit - manchmal kann zwar eine Einbindung der Daten erfolgen, aber Vokabulare dürfen nicht angereichert oder abgewandelt werden oder solche Erweiterungen und Derivation dürfen abschließend nicht mit anderen geteilt werden
    - auch die FAIRness relevanter Ressourcen wurde bei ihrer Erstellung noch nicht bedacht - einfach, weil es zum Zeitpunkt ihrer Entstehung noch kein Thema war (vor allem die Langzeitverfügbarkeit)
- das schöne Bild vom Puzzle, bei dem alle Teile naht- und mühelos ineinander passen, kann man in der Realität also leider nicht so einfach erreichen
- unter der Haube der FID-Services macht das manchmal viel, manchmal wenig aus > Erweiterungen für eine Freitextsuche wie eine Suchtermerweiterung oder ein Autcomplete lassen sich damit trotzdem implementieren, solange man unterschiedliche Ressourcen aufeinander abbildet und Crosskonkordanzen erstellt
- manche FID-Dienste setzen dagegen auf die visuelle Navigation eines Begriffssystems - hierfür wird dann eher eine einzelne Ressource mit einer überschaubaren Zahl von Einstiegsknoten (idealerweise mindestens in deutscher und englischer Sprache!) bevorzugt
    - eine solche Ressource erhält man mit dem modularen Ansatz aber nicht - jedes Vokabular kann seine eigene Begriffshierarchie postulieren, die nicht unbedingt mit der anderer Vokabulare kompatibel ist
- auch für ein Vorschlagswesen fehlt in einem multimodularen System der “Zielort” und eine gemeinsame Vokabularentwicklung mit der Fachcommunity ist schwierig - ideal wäre ein zentrales Vokabular, das Vorschläge aus der deutschen Forschungscommunity aufnehmen könnte - hier spricht natürlich viel für die GND, da sie zentrales verbales Erschließungsinstrument für die Bestände im FID move als auch im FID Bau ist

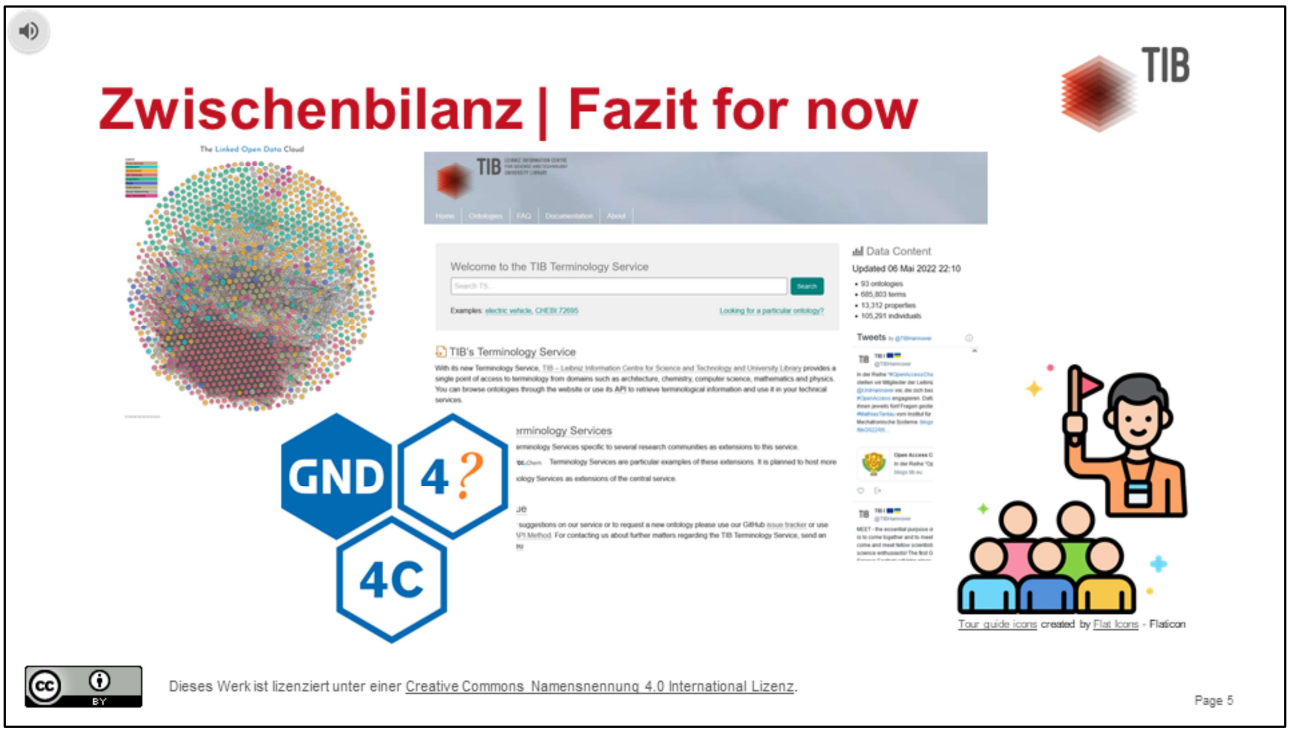



- kommen wir jetzt zu den Erschließungsszenarien in den FIDs move und BAU
  - die zu erschließenden Bestände sind ebenfalls heterogen - neben die traditionellen Publikationen werden zunehmend auch andere Medientypen fokussiert, besonders prominent z.B. Forschungsdatensätze (also Sammlungen von Dateien, die in Forschungsprozessen erhoben wurden und im besten Fall nach gängigen Metadatenstandards erschlossen wurden), aber auch Personendatensätze von Forschenden
  - die Sacherschließung von Publikationen ist traditionell an die Fachreferate gebunden und erfolgt in den etablierten Erschließungssystemen
    - hier wäre z.B. die WinIBW zu nennen
    - der Erschließungsworkflow ist dabei auf bestimmte Vokabulare festgelegt (z.B. GND und BK)
    - die Nutzung von Fachthesauri ist in diesem Workflow nicht vorgesehen
    - die Einbringung von Fachterminologie müsste über Neuansetzungen erfolgen, was aber teilweise durch das Regelwerk selbst geblockt wird:

        - Zerlegungskontrolle richtet sich gegen Komposita
        - Spezialbestände der FIDs bieten außerhalb des FID keine Ansetzungsgrundlage
        - außerdem soll ja auch nicht alles aus dem Fachthesaurus in den Allgemeinthesaurus wandern
- für Forschungs- und Personendaten finden sich dagegen unterschiedliche Ansätze:
    - solche Daten erhalten teilweise eigene Publikationssysteme, die auf Open Source-Softwares wie CKAN oder Vivo basieren
    - die Erschließungsszenarien unterscheiden sich hier:
        - teils wird basierend auf vorhandenen Erschließungselementen eine Übertragung auf den neuen Erschließungsgegenstand vorgenommen (z.B. Publikation zu Person)
        - teils wird automatisiert nacherschlossen - was teilweise - je nach eingesetztem Verfahren - auch zu kuriosen Ergebnissen führen kann
        - bei eine Zusammenführung von heterogen erschlossenen Beständen spielen auch Crosskonkordanzen und Mappings eine wichtige Rolle, um Erschließungslemente auf den gesamten Bestand zu übertragen
    - teilweise bringen die Publikationssysteme auch neue Möglichkeiten zur Erschließung mit, die von den Nutzern selbst bedient werden können
        - in VIVO-basierten Systemen ist generell die Einbindung mehrerer Vokabulare möglich, sofern sie das richtige Format haben
            - für die Nutzenden kann das mitunter Frustpotential haben, z.B.
                - wenn nicht klar ist, welches Vokabular man wählen sollte
                - wenn die Vokabulare inhaltliche Überschneidungen haben
                - wenn fachliche Konzepte fehlen
            - auch für die Betreiber kann das unschön sein - insbesondere, wenn die Anwendung die Navigation über die Begriffsrelationen

ermöglichen soll

- in CKAN-basierten Systemen kann dagegen auch eine Verschlagwortung mit freien Schlagwörtern erfolgen
    - ein Alptraum für die Standardisierung!
    - aber: dies könnte auch potentieller Input für die Erweiterung kontrollierter Vokabulare sein, der durch die FIDs ausgewertet werden könnte
    - fraglich ist aber auch, inwiefern die Nutzer solcher Ressourcen überhaupt zu einer ausführlichen Metadatenanreicherung zu bewegen sind…




Ich würde jetzt gern eine Zwischenbilanz ziehen bzw. ein vorläufiges Fazit geben, an das wir in der Diskussion hoffentlich noch einmal anknüpfen können

- zunächst einmal möchte ich festhalten, dass es eine Vielfalt an Vokabularen gibt
- diese Vielfalt sollten wir anerkennen, indem wir ihre Nutzung besser unterstützen
    - das können wir z.B. erreichen, indem wir flexiblere Erschließungsprozesse und Erschließungsumgebungen erstellen
    - eine andere wichtige Voraussetzung sind gut sortierte Terminologieservices, die die Vokabulare in den Erschließungsumgebungen bereitstellen können
- trotzdem sollten wir eine Öffnung der Big Player als zentrale Anlaufstellen für Neuansetzungen dennoch weiter im Auge behalten und weiter vorantreiben - parallel zur Community-Arbeit

Was fehlt jetzt aber noch, um mit der Vielfalt der Vokabulare besser umgehen zu können?

- hier sehe ich vor allem die Einhaltung von Mindeststandards
    - einmal bezüglich der FAIRness der Vokabulare
    - aber auch Möglichkeiten die Akzeptanz dieser Vokabulare in den Fach-Communities zu erfassen sowie ihre Qualität zu bewerten
- auch eine gute Vernetzung der Vokabulare untereinander ist immer essentiell, wenn man mit mehreren Vokabularen arbeiten möchte

- auch Möglichkeiten zur Visualisierung gemappter Vokabulare und ihrer Versionen fehlen meines Erachtens noch
- um Fachcommunities zur Weiterentwicklung existierender Vokabulare bzw. zur Erstellung neuer Spezialvokabulare zu befähigen, wäre zudem eine dem Data Stward vergleichbare Rolle - der Terminologie-Lotse - sinnvoll, der Domänenexperten, den Weg durch die Vokabularwelt zeigen kann


LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY
TIB
Danke!
Haben Sie Fragen?
Werkstattbericht
FID move ~ FID BAUdigital
Mit heterogenen Quellen…
Susanne Arndt, M.A.
Technische Informationsbibliothek Hannover
01.06.2022, Leipzig
Member of Leibniz Association
Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.