# Using multiple Indices with VuFind

Oliver Goldschmidt
TU Hamburg-Harburg
Germany

VuFind Summit 2012
Villanova University
October 16, 2012

TUHH
Technische Universität Hamburg Harburg

## Who is the guy presenting that stuff?

- working at the University Library of the University of Technology in Hamburg, Germany
  - developing library applications
  - since two and a half years mainly Vufind-based stuff
  - administrating servers
- Interested in search engines and search technology

*Photo by Heiko Weier*

# Librai

TUHH
Technische Universität Hamburg Harburg

- TUHH is member of the Common Library Network GBV (light blue)
- The Common Library Network has a service center called VZG (Verbundzentrale)
- University Library Leipzig (also attending here) is member of the SWB (South Western Library Network, red)
- SWB has a service center called BSZ (Library Service Center)
- VuFind is getting stronger in Germany

Hamburg-Harburg  Konstanz (BSZ)  Leipzig  Göttingen (VZG)

Other VuFind users

# Whats

TUHH
Technische Universität Hamburg Harburg

- You can split one index into several parts – these are the shards
  - Split your VERY LARGE index into several smaller ones
  - Index very different types of material easily and distinguish them by index
- You have more than one index
- All shards can be queried at once
- Shards can run on the same machine (but they do not need to), using a multicore environment

*Photo by Paul Atherton via Wikimedia Commons (http://commons.wikimedia.org/wiki/File:Prehistoric_pottery_shards,_Sierra_Leone.jpg)*

# Using

TUHH
Technische Universität Hamburg Harburg

**Solr can do that! VuFind also can do that!**

**Doing it, you shall keep in mind:**

- Indices should be structured similarly
- Indices need to have distinct IDs

*Photo by David Shankbone via Wikimedia Commons (http://commons.wikimedia.org/wiki/File:Photovoltaic_arrays_at_the_Israeli_National_Solar_Energy_Center.jpg)*

BUT: Reality might be different. What if you have indices, which are not matching these requirements?

# Shard

TUHH
Technische Universität Hamburg Harburg

| Index A | Index B |
|---|---|
| author | author |
| collection | n/a |
| recordtype | recordtype |
| title | title |

**What if my indices are not structured completely similar?**

**MINDEX**

- Your shards are structured differently
- You still want to use sharding to search all indices at once
- Solution: For fields, which are not present in ALL of the cores you have to do „Field Stripping“
- Field Stripping might cause hardly explainable results
- Field Stripping is not necessary in Dismax Searchhandler, but for Lucene Searchhandler
- It's still not adviced to create a MINDEX situation! You better add missing fields to each index!

TUHH
Technische Universität Hamburg Harburg

## What if my indices have similar IDs?

Result for *:
1. VuFind
2. VuFind2
3. Developing

| Shard 1 ID | Shard 1 title | Shard 2 ID | Shard 2 title |
|---|---|---|---|
| 1 | VuFind | 1 | VuFind |
| 2 | VuFind2 | 2 | VuFind2 |
| | | 3 | Developing |

Result for * option 1:
1. VuFind
2. VuFind2
3. Developing

Result for * option 2:
1. VuFind
2. VuFind 2.0
3. Developing

| Shard 1 ID | Shard 1 title | Shard 2 ID | Shard 2 title |
|---|---|---|---|
| 1 | VuFind | 1 | VuFind |
| 2 | VuFind2 | 2 | VuFind 2.0 |
| | | 3 | Developing |

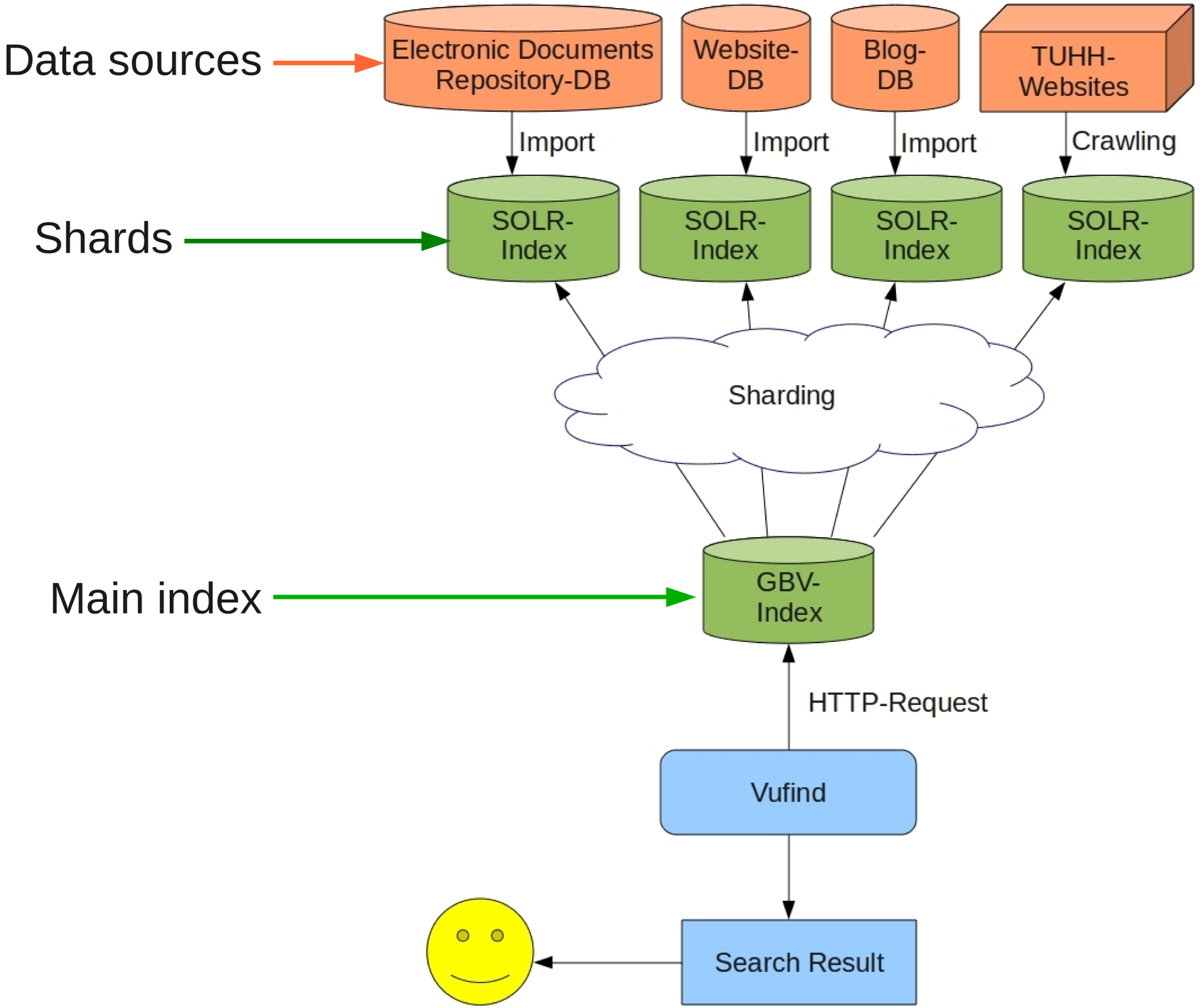
Result for *:
1. VuFind
2. VuFind
3. VuFind2
4. VuFind 2.0
5. Developing

| Shard 1 ID | Shard 1 title | Shard 2 ID | Shard 2 title |
|---|---|---|---|
| 1 | VuFind | web-1 | VuFind |
| 2 | VuFind2 | web-2 | VuFind 2.0 |
| | | web-3 | Developing |

Though its still not adviced to use similar IDs, even if they had the same data behind them (data may change)! It might be better to use prefixes!

# How we use shards

TUHH
Technische Universität Hamburg Harburg

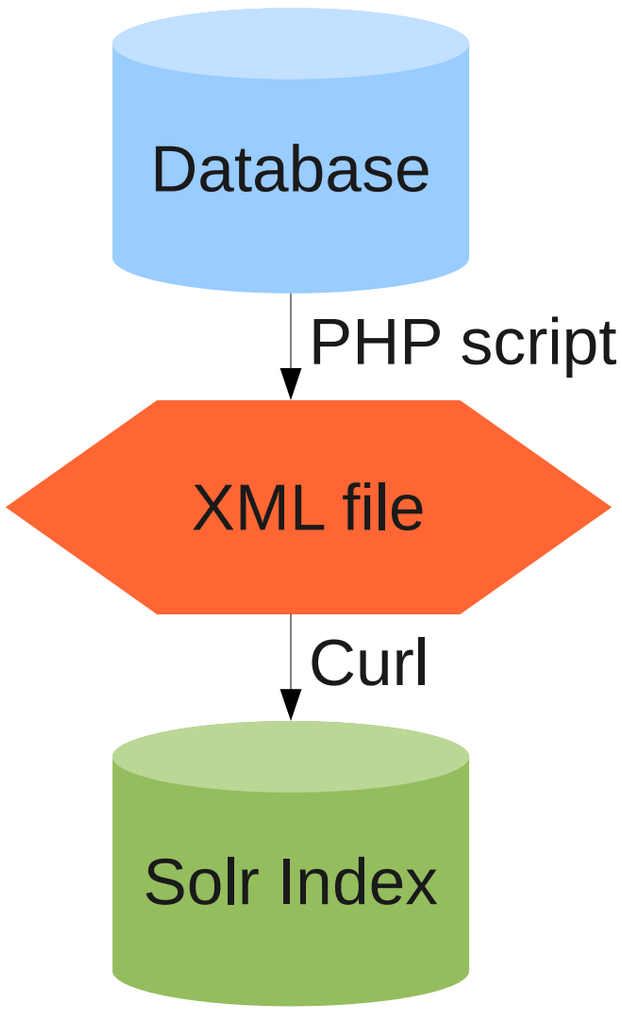


# How v

TUHH
Technische Universität Hamburg Harburg

For the Website-index we do not have MARC-data.

We have the data structured in a database.

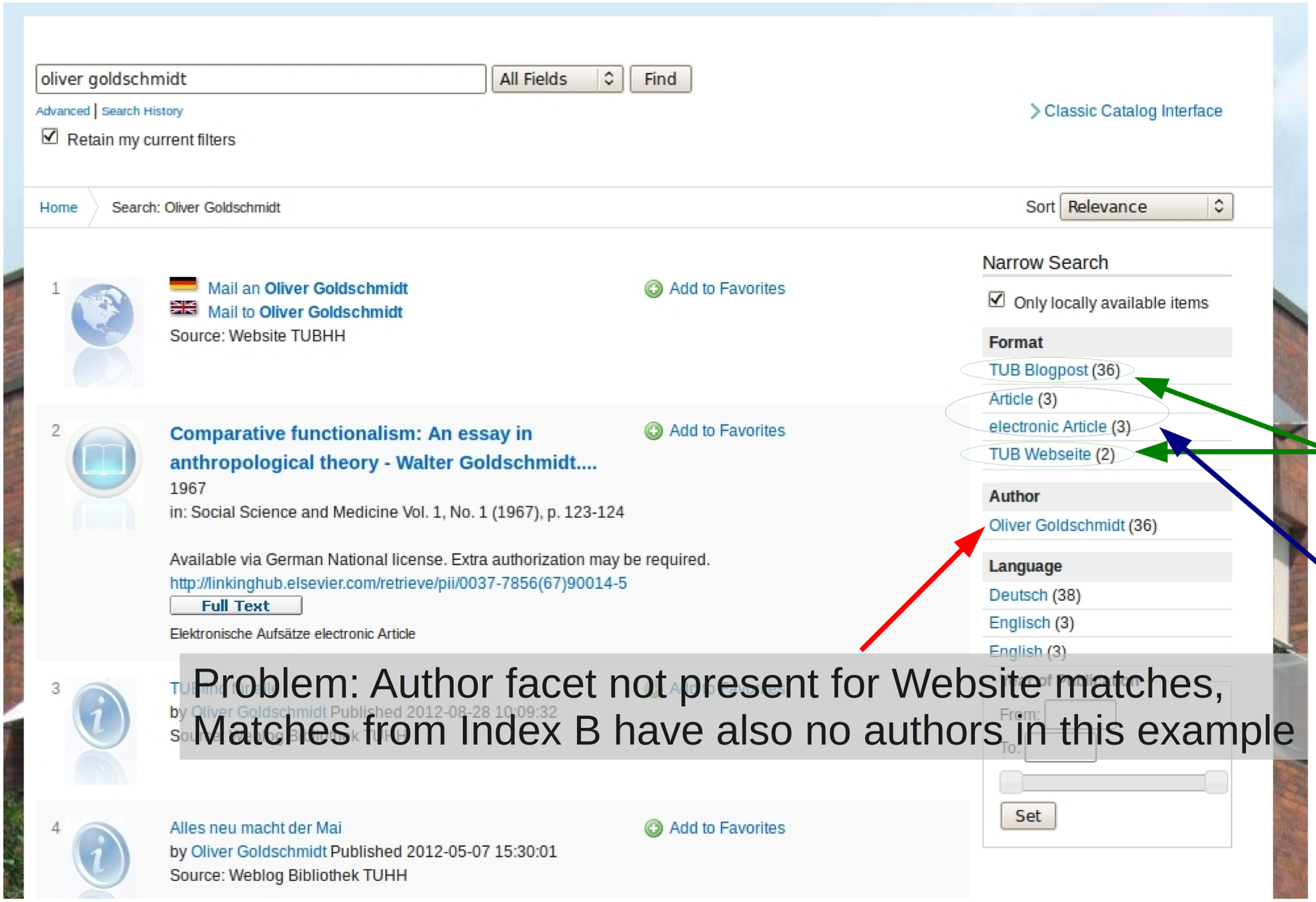



```
<?xml version="1.0" encoding="UTF-8"?>
<add>
    <doc>
        <field name="institution">TUHH-Bibliothek</field>
        <field name="location">Internet</field>
        <field name="recordtype">WPsite</field>
        <field name="collection">Website</field>
        <field name="format">TUB Webseite</field>
        <field name="id"><![CDATA[website-19]]></field>
        <field name="publishDate"><![CDATA[2012-05-11 14:18:15]]></field>
        <field name="url"><![CDATA[http://www.tub.tu-harburg.de/?p=19]]></field>
        <field name="contents"><![CDATA[Discover your library resources:
        Or search our holdings the classic way:
        Which search should I use?]]></field>
        <field name="titleEng"><![CDATA[Find]]></field>
        <field name="title"><![CDATA[Find]]></field>
        <field name="language">Englisch</field>
        <field name="url"><![CDATA[http://www.tub.tu-harburg.de/en/?p=19]]></field>
    </doc>
</add>
```

```
/usr/bin/curl -d @xmlfile.xml \
http://solrhost/solr/index/update \
-H "Content-Type: text/xml; charset: UTF-8"
```

# Prese

TUHH
Technische Universität Hamburg Harburg




Problem: Author facet not present for Website matches,
Matches from Index B have also no authors in this example

Multiple indices are queried at once. The result is shown in one single result list.

Index A (Weblog and Website)

Index B (Index from Common Library Network in Germany)

# Scena

TUHH
Technische Universität Hamburg Harburg

You could consider using scenario 1 in an environment, where

✔ You want to combine different indices to one result list

✔ You want to have one common facet list for this result list

✔ You want to integrate an external index from ExLibris (Primo Central)

Be warned, if

! You have different schemas (i.e. different fields) for your indices (field stripping will be needed for queries using Lucene Searchhandler)

! You have different facet names for actually equivalent values or one shard does not have facet values for certain fields at all (facets can cause confusion)

Its impossible to use Scenario 1 in an environment, where

✗ You want to integrate an external index from Summon or EBSCO (you cannot get only one result list with them)

✗ You don't have at least reading access on the Solr server of any of the sharded indices (using an API solution instead)

# Prese

TUHH
Technische Universität Hamburg Harburg

Multiple indices are available to Vufind, but they are not searched at once.

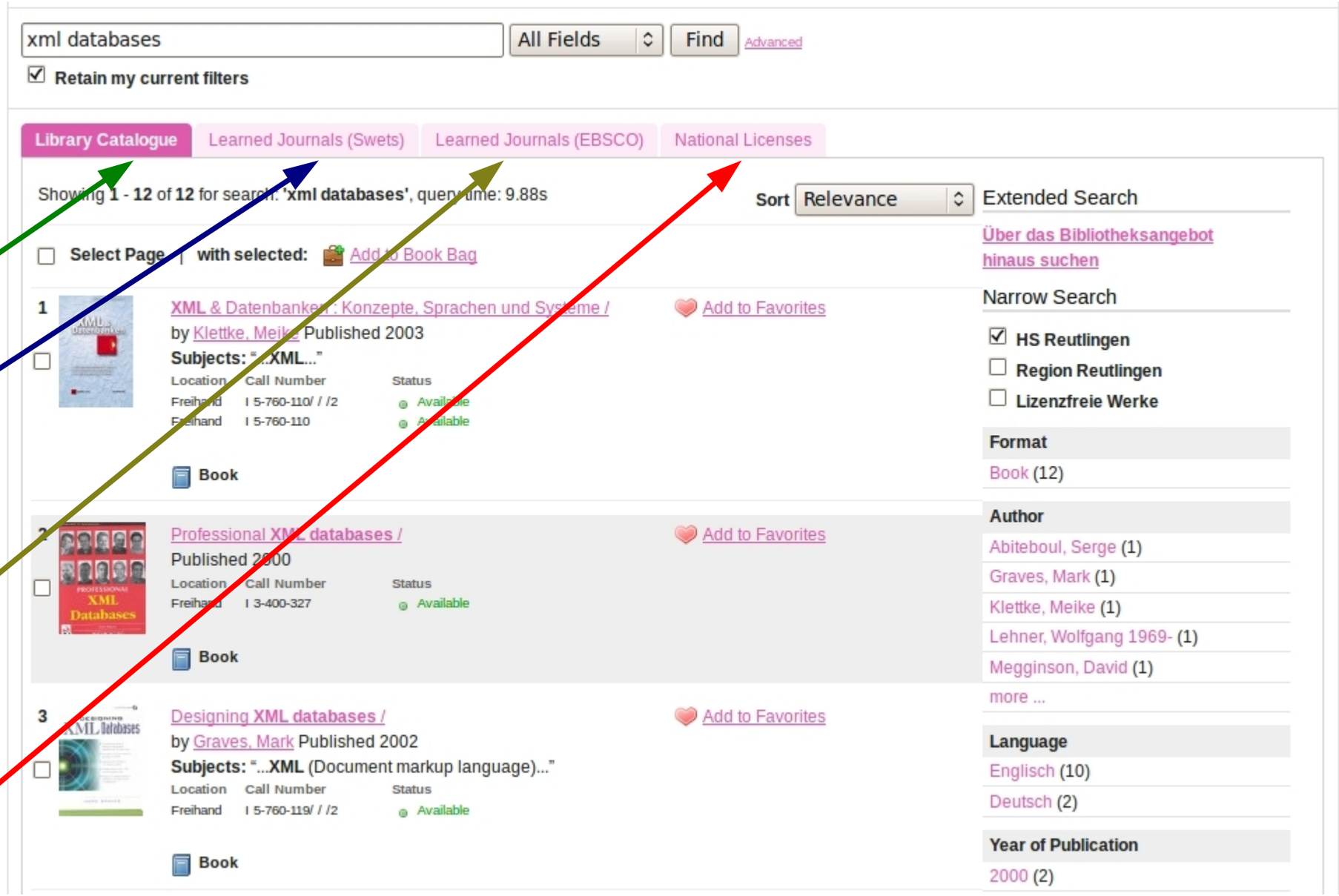
User has to select a register.

Index A

Index B

EBSCO: Special index, not available via sharding

Index C




This example is taken from BOSS (BSZ One Stop Search) from the BSZ in Germany.

# Scena

TUHH
Technische Universität Hamburg Harburg

You could consider using scenario 2 in an environment, where

✔ You want to combine different indices in one VuFind instance

✔ You want to integrate an external index from Primo Central, Summon or EBSCO

✔ You want to avoid the issues you know from scenario 1

Remember:

! You will have different result lists for any of the shards, each one having its own facetting list

! What you are doing here, is actually not real sharding, because all shards are displayed in seperate result lists

# Outlo

TUHH
Technische Universität Hamburg Harburg

- VuFind 1.x shard support could get improved (some issues discussed in Jira)
- VuFind 2 will still support using shards and will have some improvements written by David Maus and Demian Katz
- StripFields could become unnecessary (Dismax does not care about missing fields, why does Lucene?)
- Sharding could become more interesting for VuFind, when we start using a central index

*Photo by opensourceway via Flickr (http://www.flickr.com/photos/opensourceway/5319988695/)*

**Thank you for your attention**

**Oliver Goldschmidt** **o.goldschmidt@tuhh.de**

**University of Technology Hamburg**
**University Library**