

# Connectors - Enablers for access to heterogeneous data sources

Dennis Jankowski

4<sup>th</sup> International Open Search Symposium

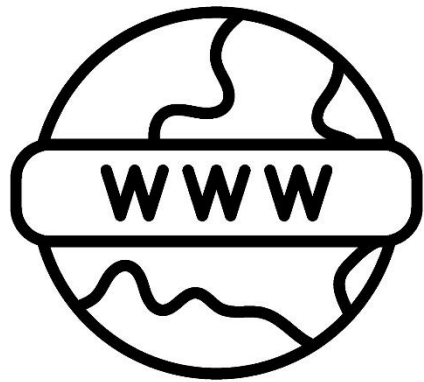
Monday, 03 October 2022



Knowledge for Tomorrow



# Introduction - Search and find any data



A white rectangular box with a black border containing several logos. At the top left is the 'Navigators' logo with a location pin icon. Below it is the 'elib' logo with the text 'Publikationen des DLR'. To the right is the 'DuckDuckGo' logo featuring a duck. At the bottom is the 'metaGer' logo in orange and black, followed by three dots.

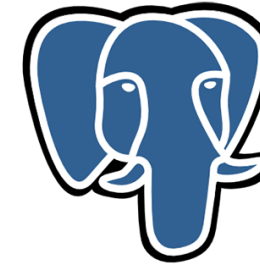


A collage of images and text. At the top left is a snippet of an article titled 'Google Dominates Thanks to an Unrivaled View of the Web'. Below it is the 'open search foundation' logo and a banner that reads 'Let's create a better way to search the internet'. To the right is a cartoon illustration of a signpost with arrows pointing in different directions, labeled 'OPEN', 'FREE', 'NEUTRAL', 'DIVERSE', and 'HAIL'. At the bottom is a snippet of a conference paper titled 'Searching on Heterogeneous and Decentralized Data: A Short Review'.



## Introduction - Search and find any data

- **Beside web data - other data sources can also provide valuable information**
  - Search should not only focus on classic web data
  - Data can also be found and retrieved in databases, REST APIs, FTP servers , ...
- **But - Wide variety of different interfaces and formats**
  - Data is queried in different ways depending on the type of data source
  - More difficult to make the data searchable for users
- **Challenge:** *How can technically heterogeneous data sources be uniformly accessed and searched?*



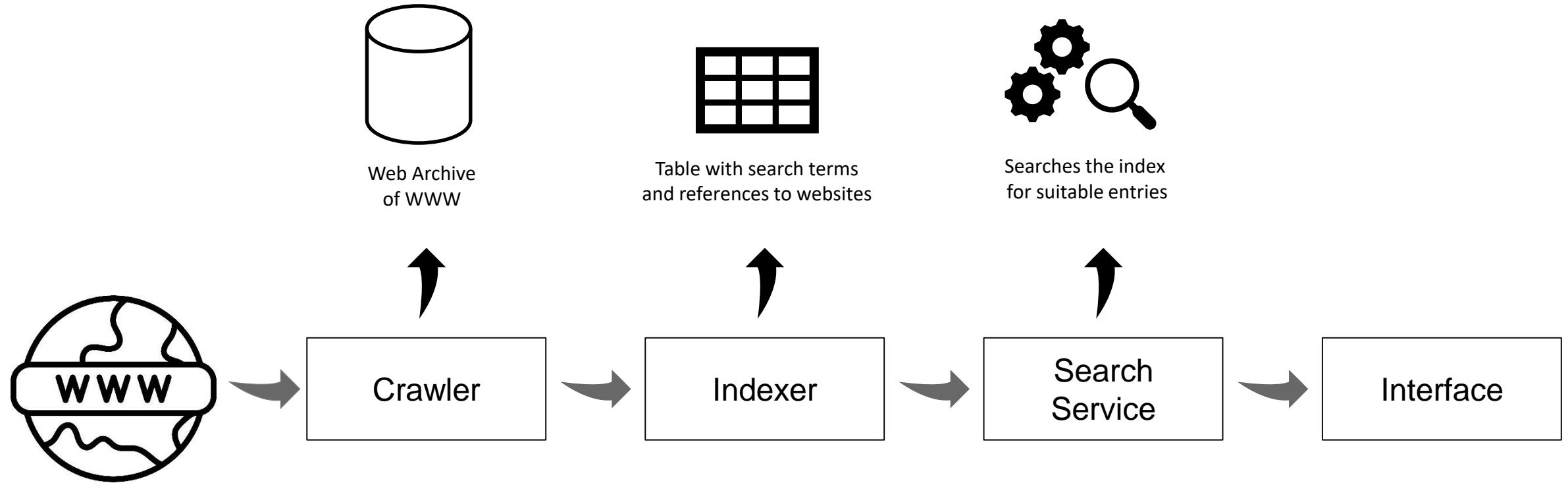
PostgreSQL



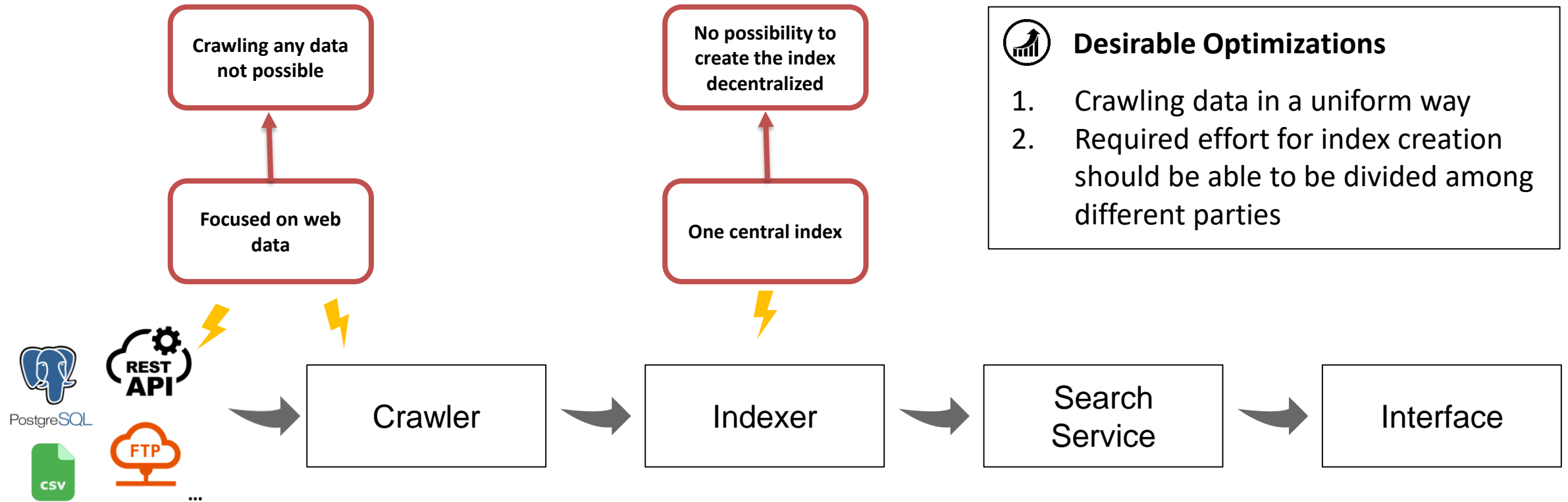
...



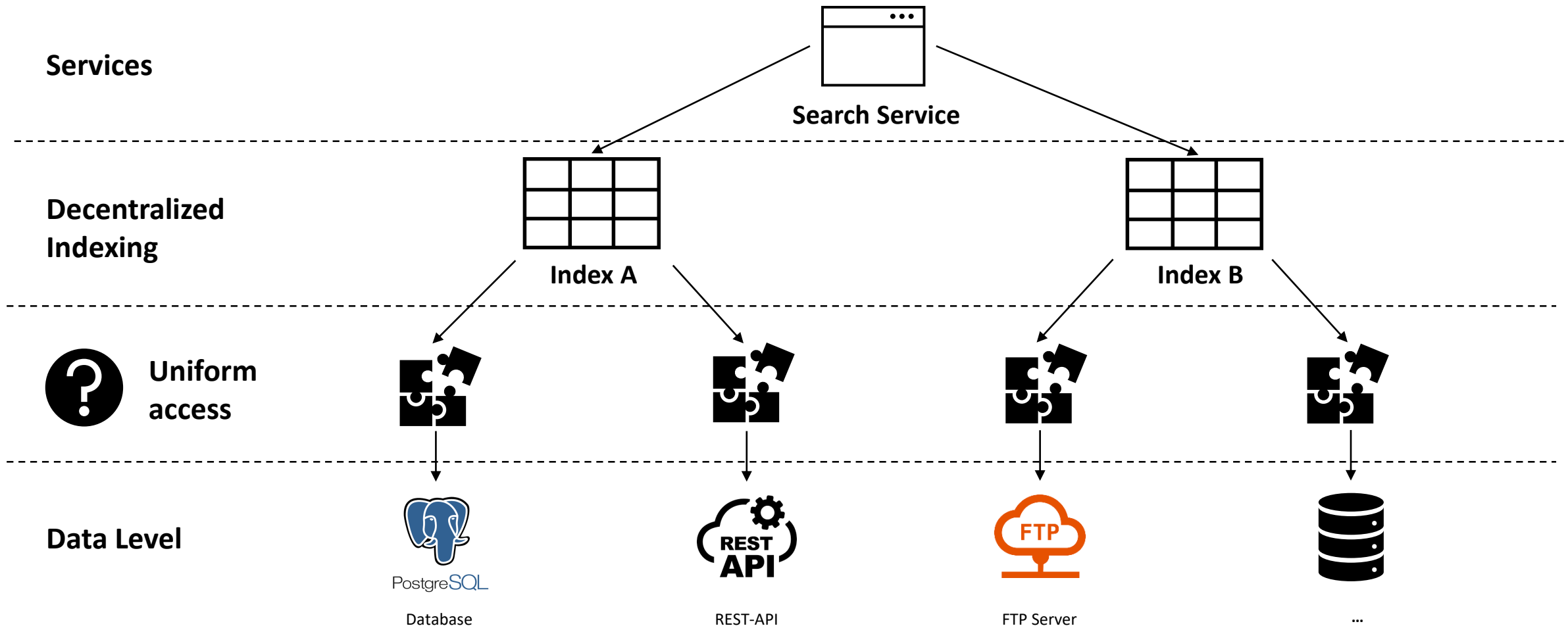
# Challenge - Structure of common Web Search Architectures



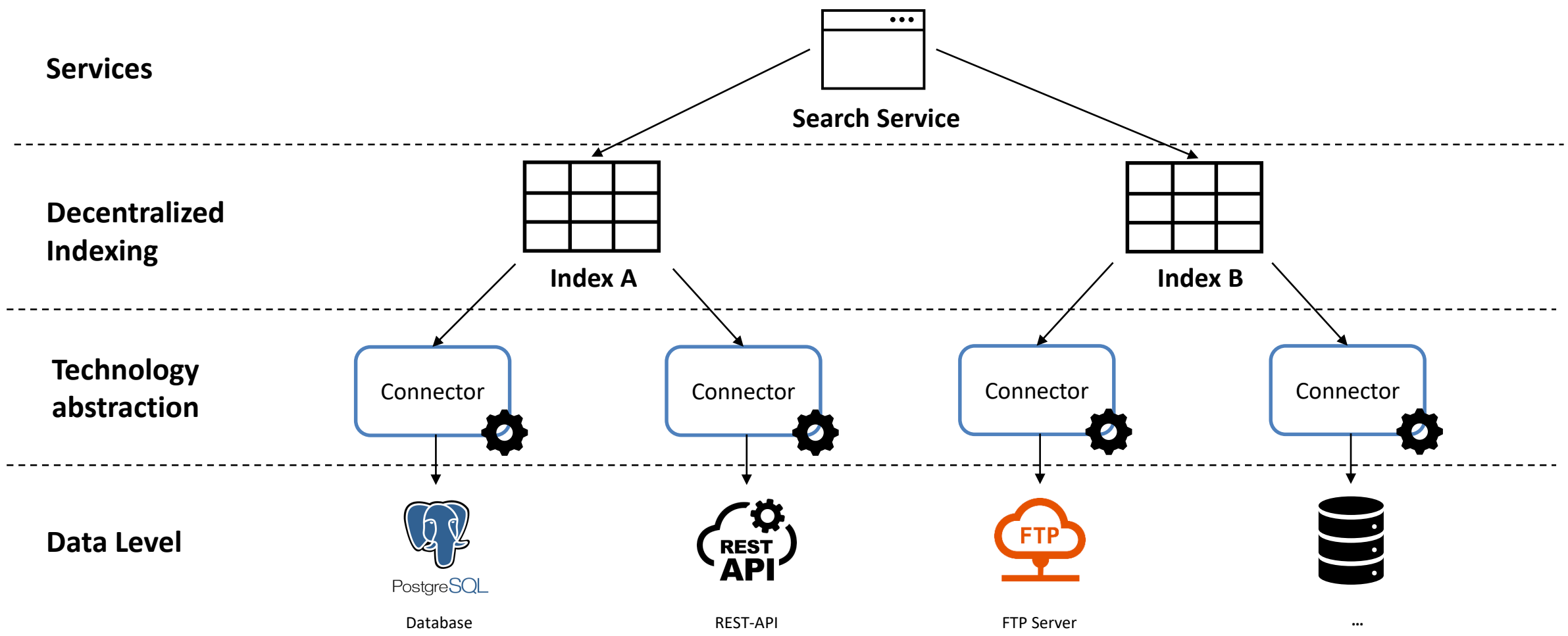
# Challenge - Limits of common Web Search Architectures



# Challenge - Architecture for Indexing Heterogenous Data Sources

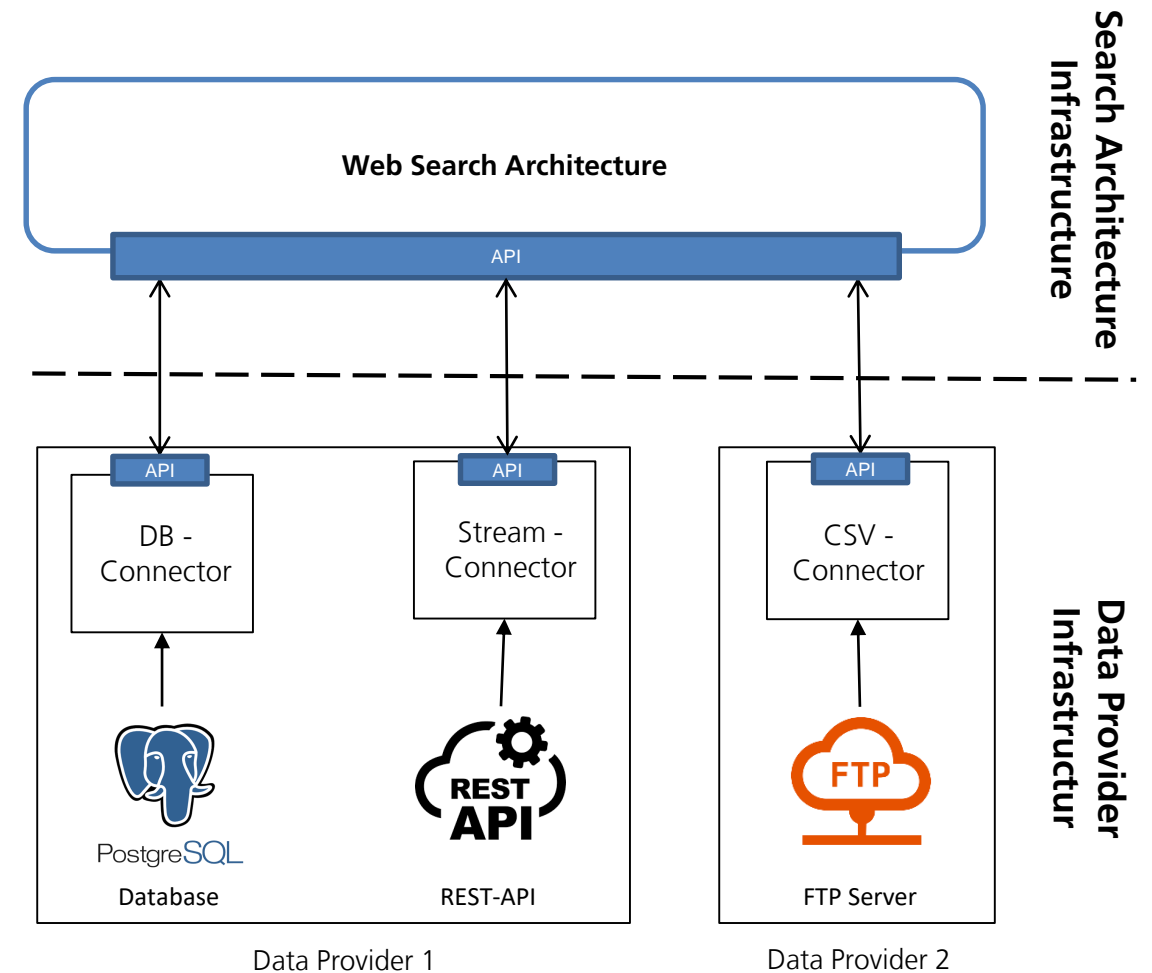


# Challenge - Architecture for Indexing Heterogenous Data Sources



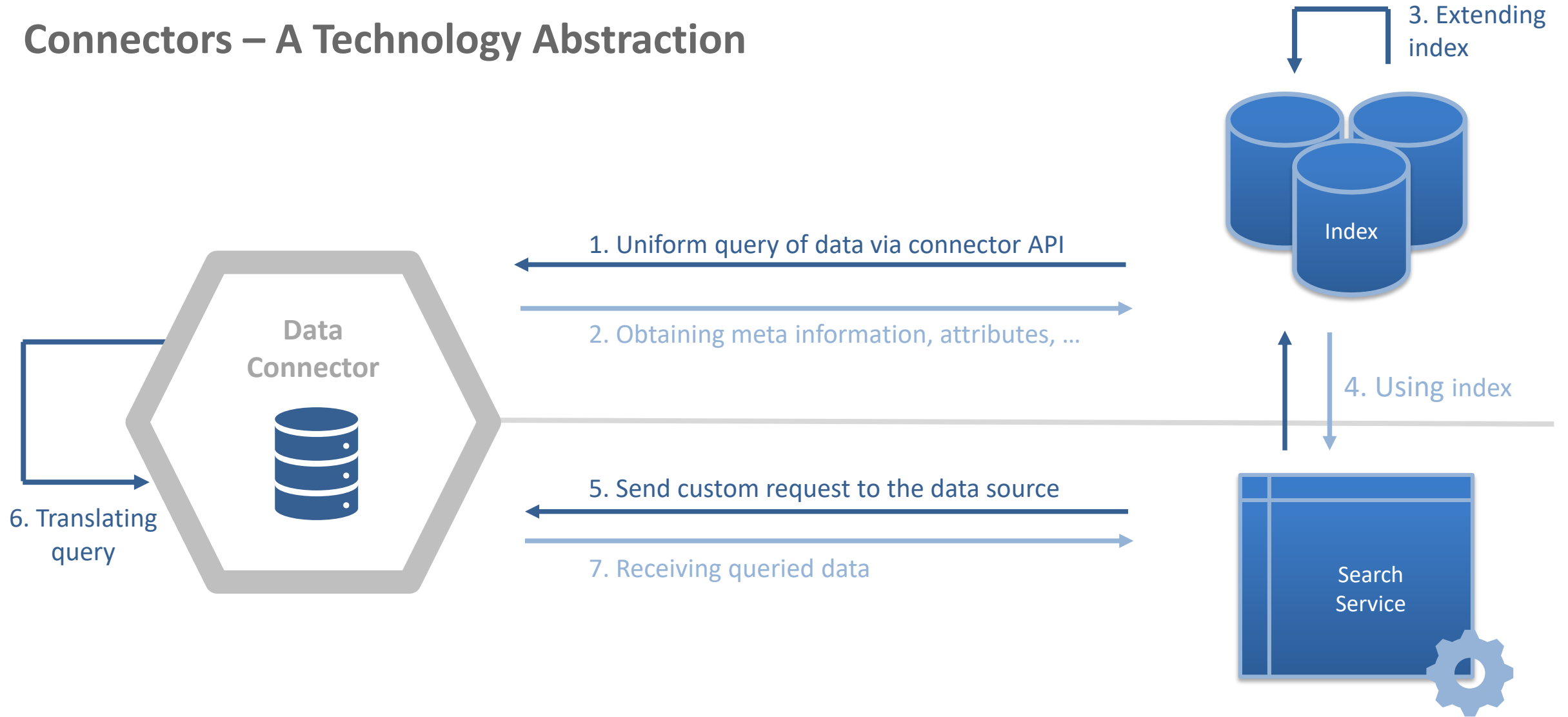
# Connectors – A Technology Abstraction

- **Uniform data access to distributed heterogeneous data sources by using "connectors"**
- **Each data source has its own connector**
- **Intermediation between Data Source and Application**
  - Technology abstraction - translates language of data source into one specific common language
  - Follows a standardized interface specification and is thus fully interoperable
- **Connectors are operated by the respective data provider**
  - Protection of data sovereignty
  - Control over provided content
  - Access control by provider
  - Simple connection and disconnection of the data source possible

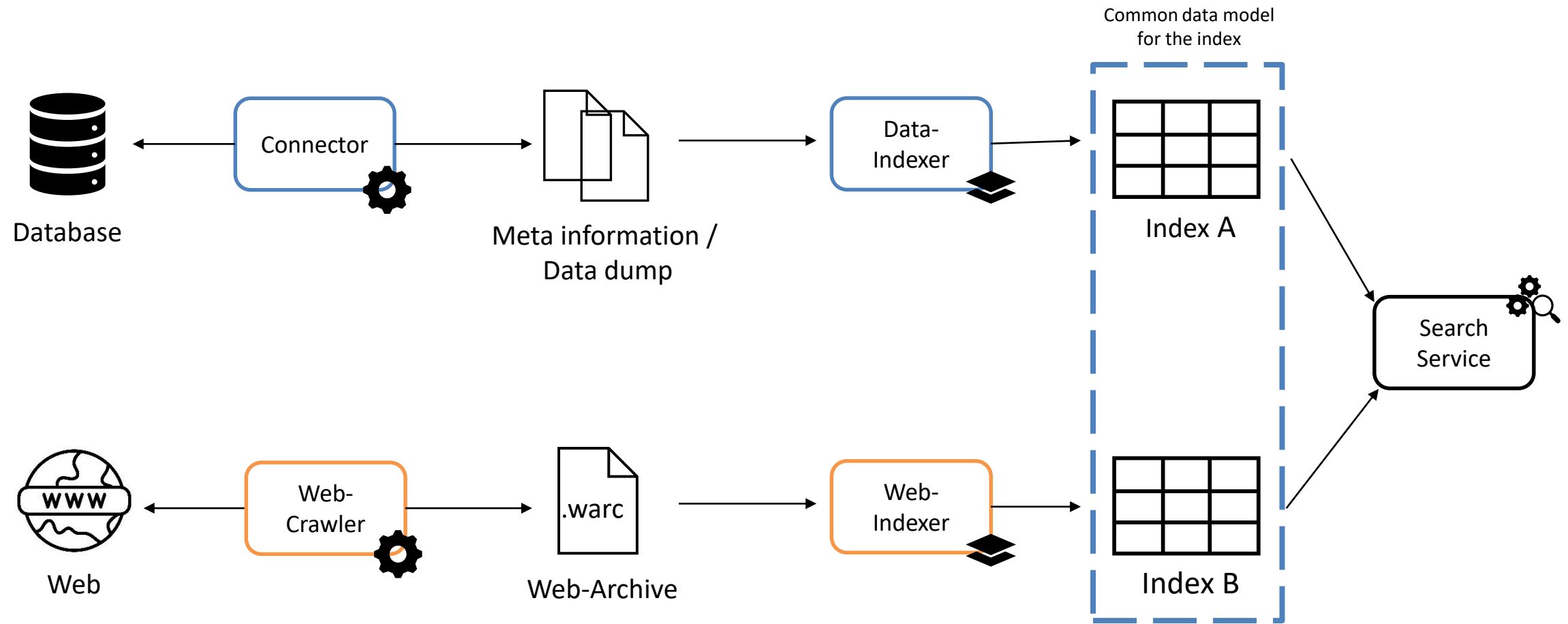




# Connectors – A Technology Abstraction



# Architecture for Indexing Heterogeneous Data Sources



# Application Example – OpenSearch@DLR Prototype

BRUNSBÜTTEL Search

---

Open Filter

AREA OF TIME

Off  Start  Start to End

Start

End

FILTER SELECTOR

Circle  Rectangle  Polygon

DATA ORIGIN

EXCLUDE CONNECTOR

POSTGRESCONNECTOR

POSTGRAPHILECONNECTOR

CSVCONNECTOR

LATITUDE: 53.9128, LONGITUDE: 9.1187

- **OpenSearch@DLR prototype**
- 3 Data Sources (Postgres Database, REST-API, CSV-File)
- Connected with 3 implemented connectors
- Heterogeneous data sources can now be searched and queried simultaneously through the UI
  - Keyword search
  - Geographic search
  - Temporal search



# Application Example – OpenSearch@DLR Prototype

## POSTGRAPHILECONNECTOR

NUMBER OF RESULTS: 497

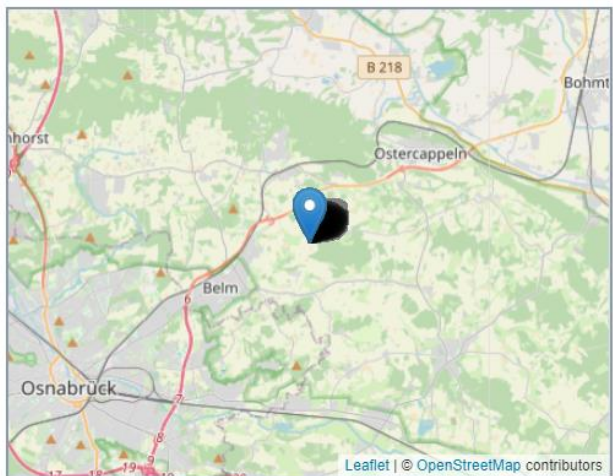
AREA OF TIME: .

START : 2016-05-30T11:06:30.000Z

END : 2020-11-05T08:00:00.000Z

CONNECTORFIELDS

```
▼"CONNECTORFIELDS" : [  
  ▼0 : {  
    ▼"NODES" : [  
      ▼0 : {  
        ▼"GEOMETRY" : [  
          0 : "GEOJSON"  
          1 : "SRID"  
          2 : "X"  
          3 : "Y"        ]  
      }  
    ]  
  }  
]
```



## POSTGRESCONNECTOR

NUMBER OF RESULTS: 2002

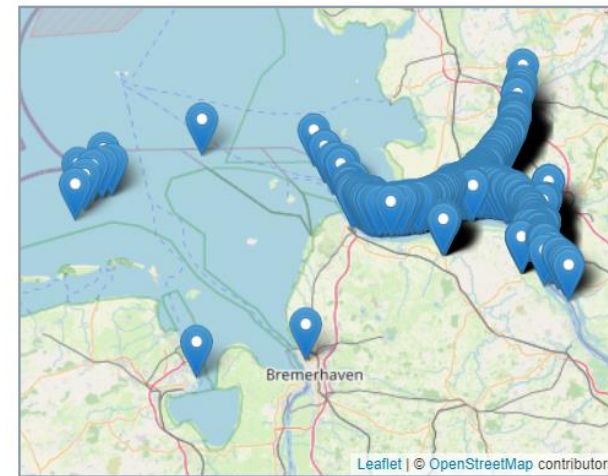
AREA OF TIME:

START : 2016-05-30T11:06:30.000Z

END : 2020-11-05T08:00:00.000Z

CONNECTORFIELDS

```
▼"CONNECTORFIELDS" : [  
  ▼0 : {  
    ▼"NODES" : [  
      ▼0 : {  
        ▼"GEOMETRY" : [  
          0 : "GEOJSON"  
          1 : "SRID"  
          2 : "X"  
          3 : "Y"        ]  
      }  
    ]  
  }  
]
```



# Summary

- Searching heterogenous data sources has **different requirements** than classic web search
  - Needs to handle different technologies and interfaces
- Connector **„speaks“** the language of data source and the common language of the service
  - Serves as a translator
  - Despite heterogeneous technologies **one uniform query language** for all data sources
- **Services** can be realized easier through uniform interface
  - Indexer / Search Service does not need to understand the language of the underlying data source

