Connector Architecture for Indexing Heterogenous Data Sources

Dennis Jankowski (DLR-SE)

WAW - OpenSearch@DLR
23.03.2022

Relevance of heterogeneous data sources

Goal - Finding the desired data

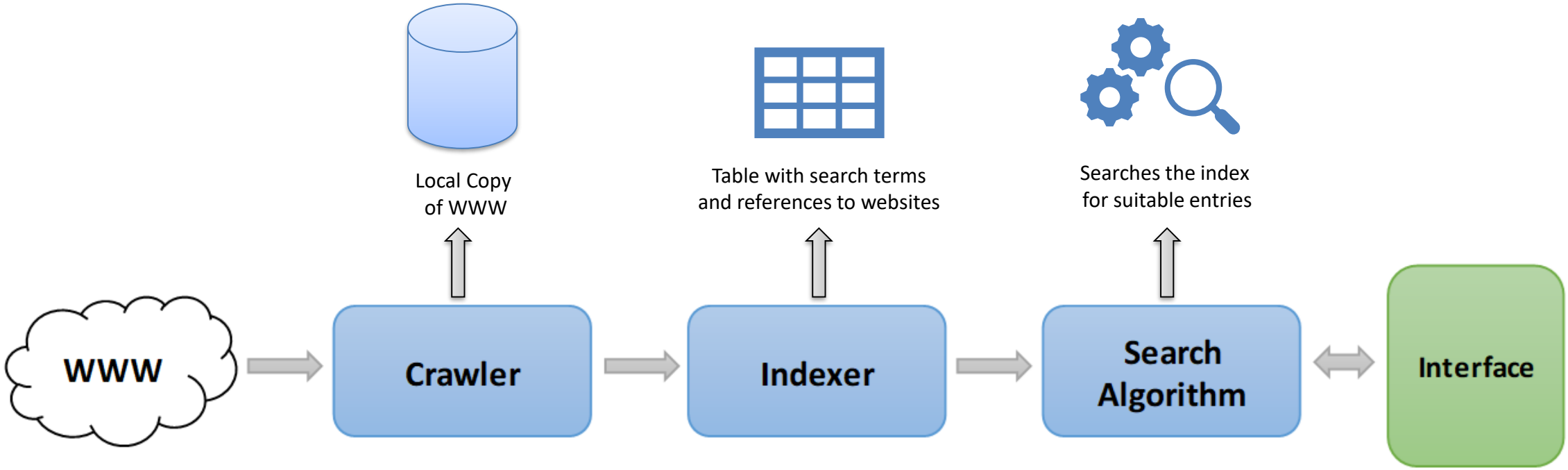
```
1  <!DOCTYPE html>
2  <html>
3      <head>
4          <title>Example<title>
5          <link rel="stylesheet" href="sty
6      </head>
7      <body>
8          <h1>
9              <a href="/">Header</a>
10         </h1>
11         <nav>
12             <a href="one/">One</a>
13             <a href="two/">Two</a>
14             <a href="three/">Three</a>
15         </nav>
```

	A	B	C	D
1	Product	Qtr 1	Qtr 2	Grand Tota
2	Chocolade	$744.60	$162.56	$907.16
3	Gummibarchen	$5,079.60	$1,249.20	$6,328.80
4	Scottish Longbreads	$1,267.50	$1,062.50	$2,330.00
5	Sir Rodney's Scones	$1,418.00	$756.00	$2,174.00
6	Tarte au sucre	$4,728.00	$4,547.92	$9,275.92
7	Chocolate Biscuits	$943.89	$349.60	$1,293.49
8	Total	$14,181.59	$8,127.78	$22,309.37

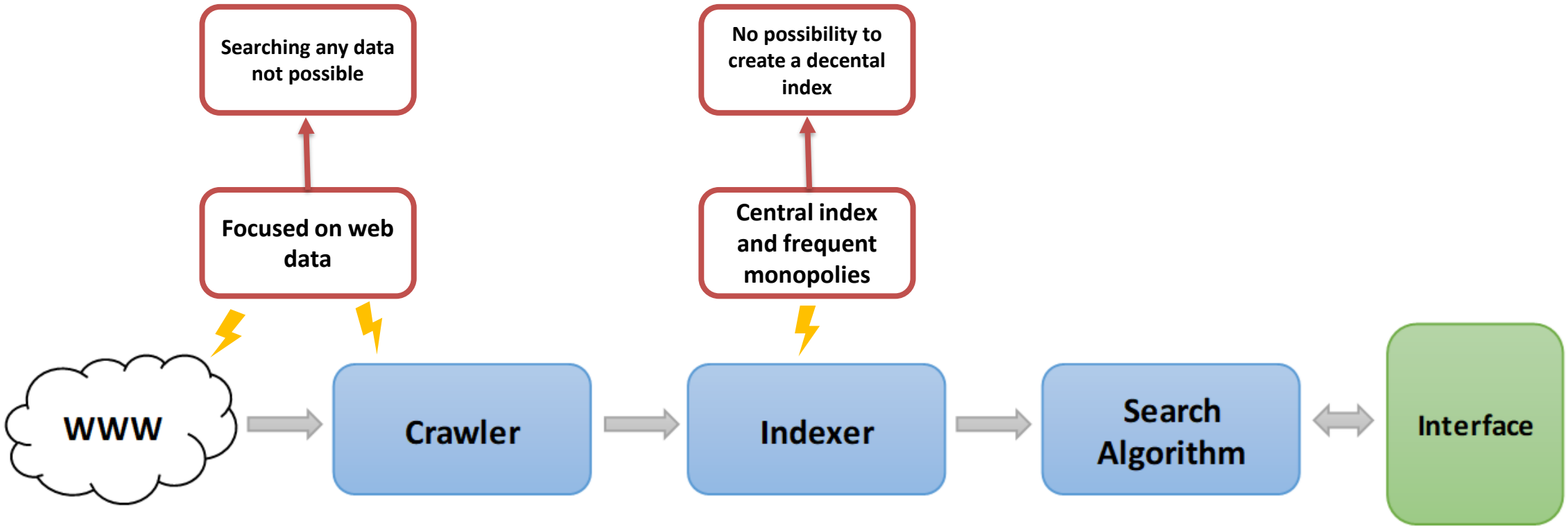
- Wide variety of formats and interfaces
  - Differs in: Content, Structure, Accessibility
- Working with data is only beneficial when relevant data can be found and accessed in a simple way

How can technical heterogeneous data sources be accessed and searched in a uniform way?

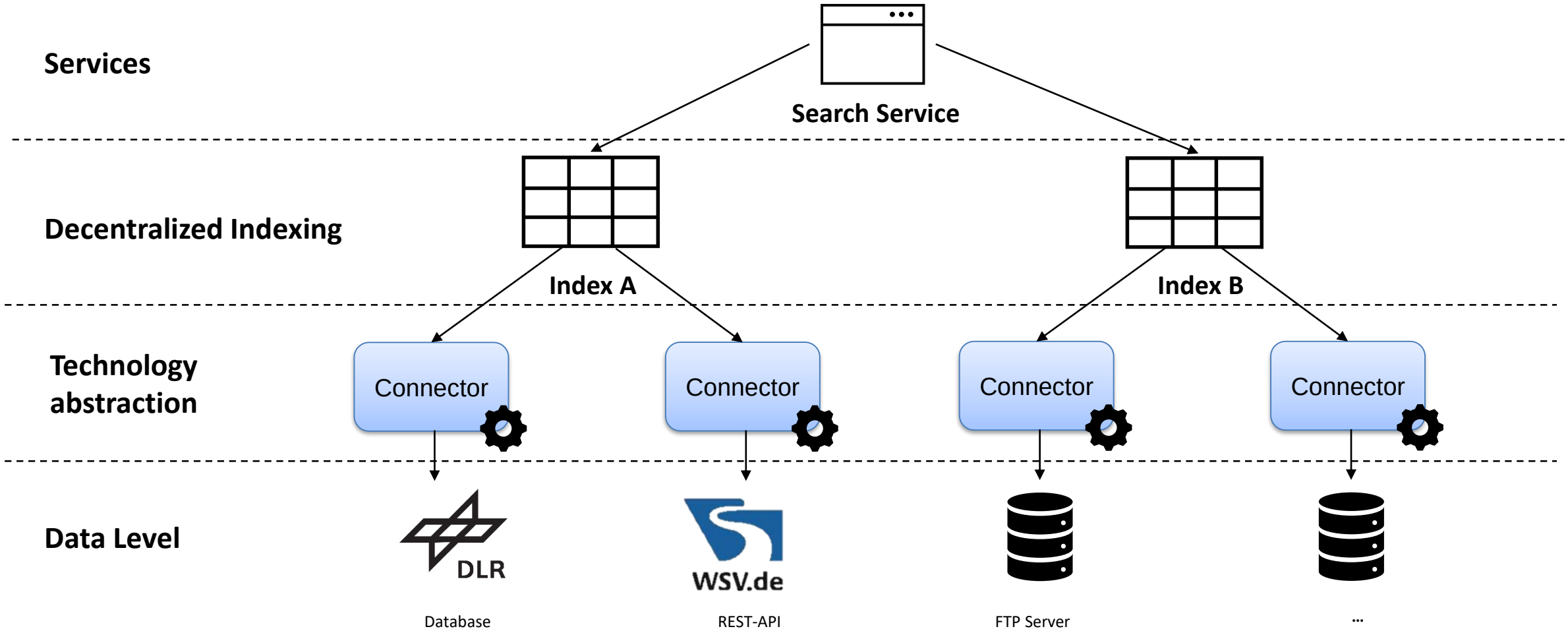
Classic Web Search Engine Architectures



Limits of Classic Search Engine Architectures



Architecture for Indexing Heterogenous Data Sources



Indexing of Non-Web Data - Connectors

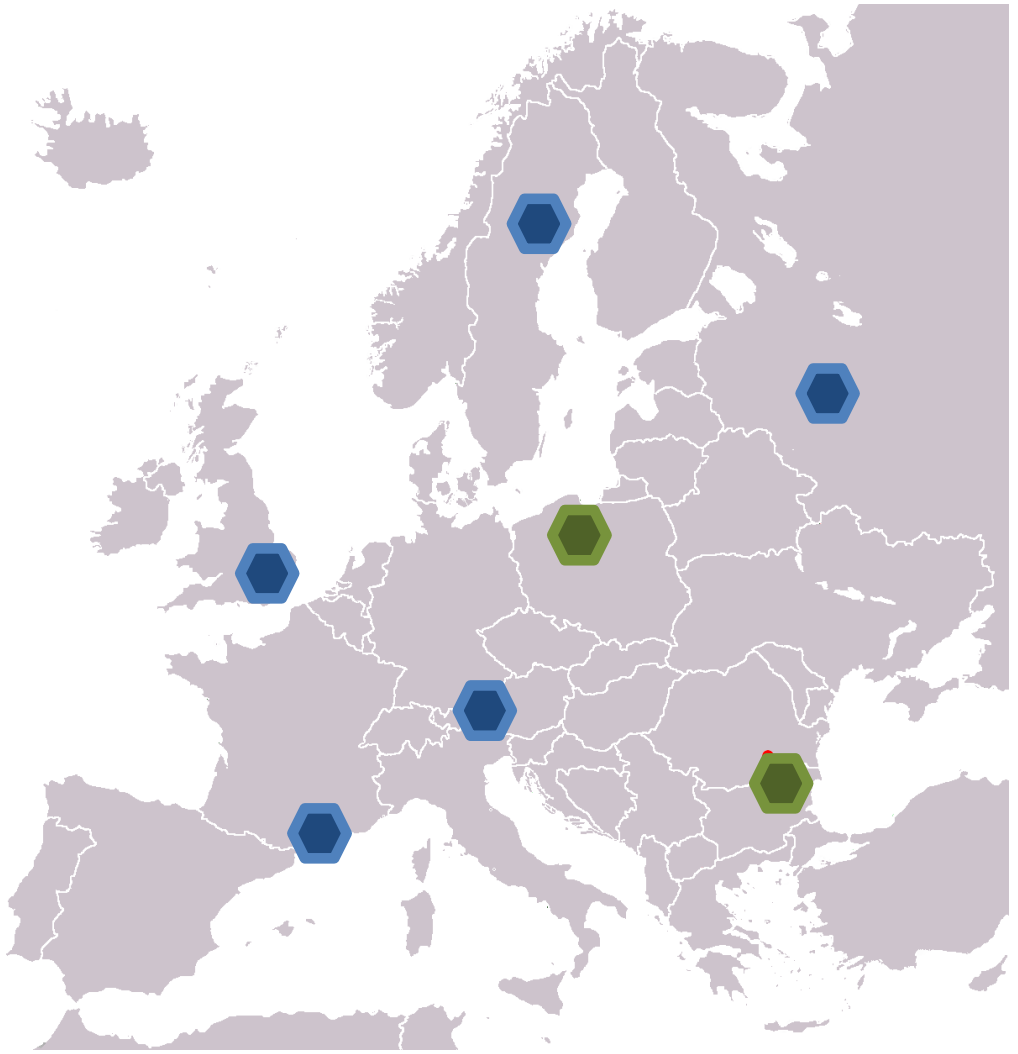
- **Each connector** (data source)...

  ... is **operated by data providers** who are interested in sharing their data

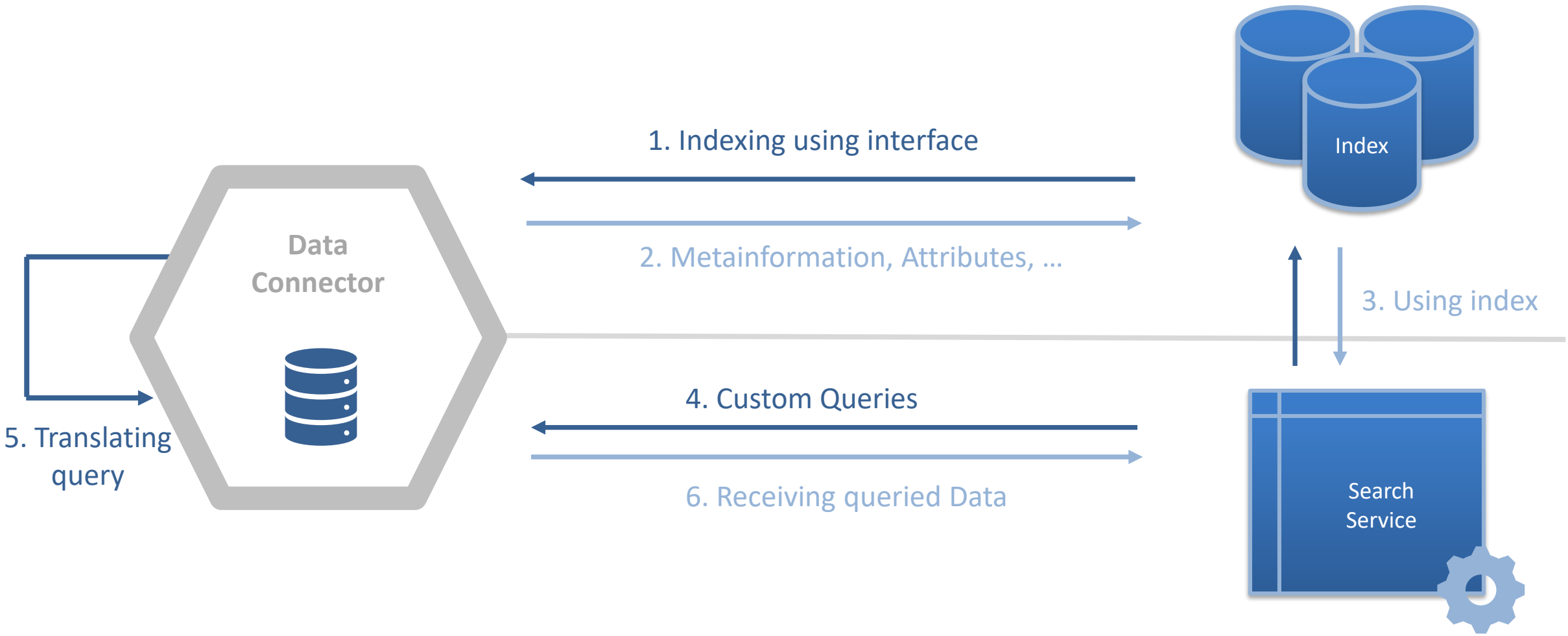
  …**control** about the published data lies by the **data provider**

  …**translates language** of data source into one specific common language

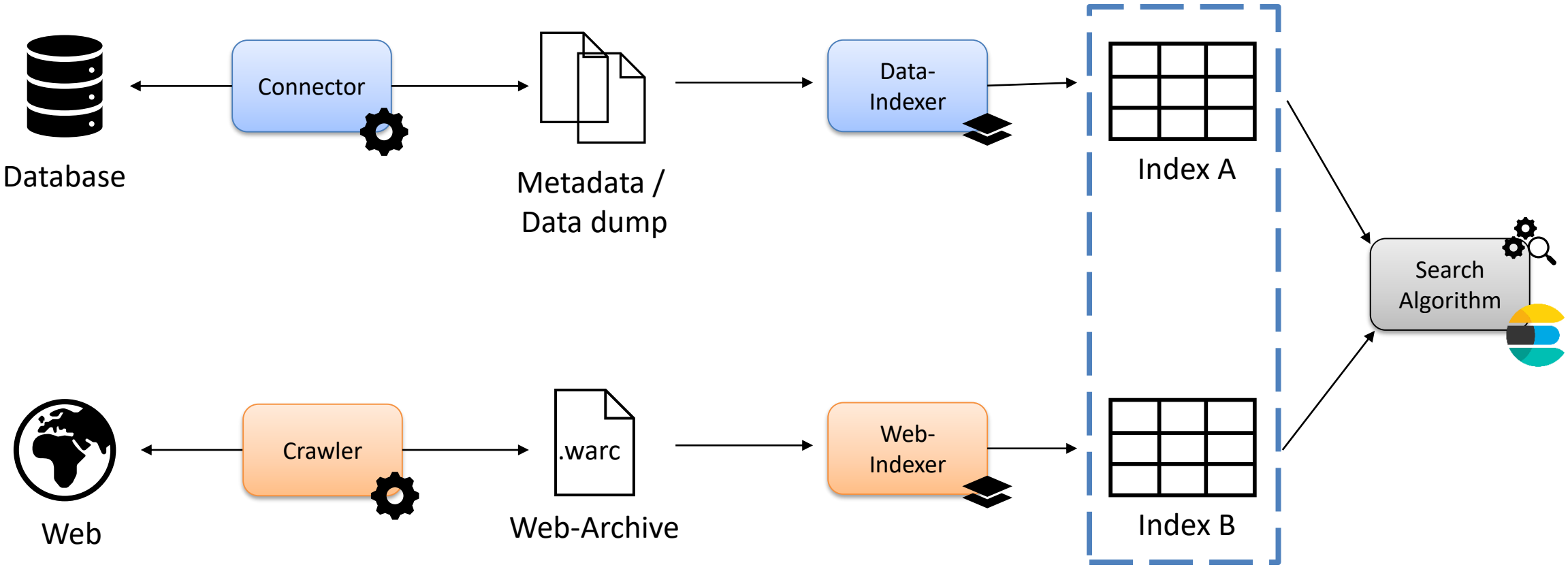
  ... follows a **standardized interface specification** and is thus fully interoperable



Indexing and Searching of Non-Web Data



Distributed Index



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