"Open Search for Science - Science for Open Search”

Stefan Voigt, German Aerospace Center, Oberpfaffenhofen, Germany
Tobias Hecking, German Aerospace Center, Köln, Germany
Dennis Jankowski, German Aerospace Center, Oldenburg, Germany
Max Schwinger, German Aerospace Center, Oberpfaffenhofen, Germany

#ossym2022, CERN, Geneva Switzerland

Science is about the discovery of new things, knowledge, concepts, contexts, systematics, relations etc.

- **The more data and information** we render and collect in the digital sphere and consume (also as the basis for science**), the more important it is to ensure open, unbiased and public access to it.**
- **Science profits significantly** from the ease of exchange of data and information and thus also has a **responsibility in keeping it open,** accessible, findable and curated.

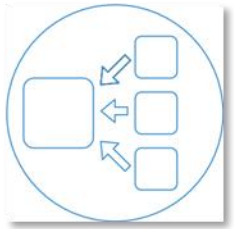
Science ⇔ Web

Large scientific organizations handle large quantities of information and data from external sources and internally in their research environments.

- **Large science organisation generate, store, manage and maintain large amounts of information** in their intranet, research systems, data repositories and information corpora as well as they generate large volumes of scientific artefacts which they put out to the public domain.

- The German Aerospace Centre, about 10.000 scientific, technical and administrative staff has a **vital interest in ensuring that access to digital information is unbiased, objective and as efficient as possible.** Internally, within the organisation, as well as externally, in the web.

Project tasks and activities



1. OpenSearch Network within DLR

- 1.1. Research on DLR internal OpenSearch potential
- 1.2. WIKI OpenSearch@DLR

2. Events and Cooperation

- 2.1. Internal Meetings
- 2.2. WAW-OpenSearch
- 2.3. Contributions to Conferences and Symposia
- 2.4. Project Cooperation and Partnerships

12-14 October 2020 FULLY WEB-BASED
2nd International Symposium on Open Search Technology

SIGIR

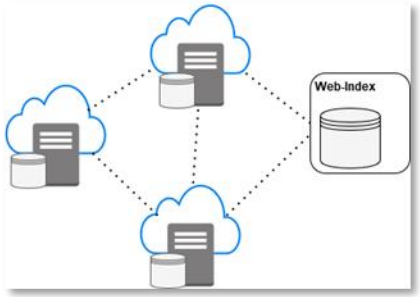
CIKM 2021 1-5 NOVEMBER

CERN

osf open search foundation

HELMHOLTZ

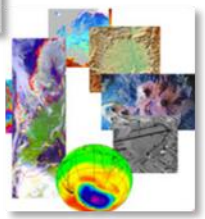
3. Pilot Applications and Project



- 3.1. DLR OpenSearch Testbed
- 3.2. Database connectivity
- 3.3. Pilot applications
- 3.4. Specialized data base connectivity

elib Publikationen des DLR

DLR.Wiki



GitLab

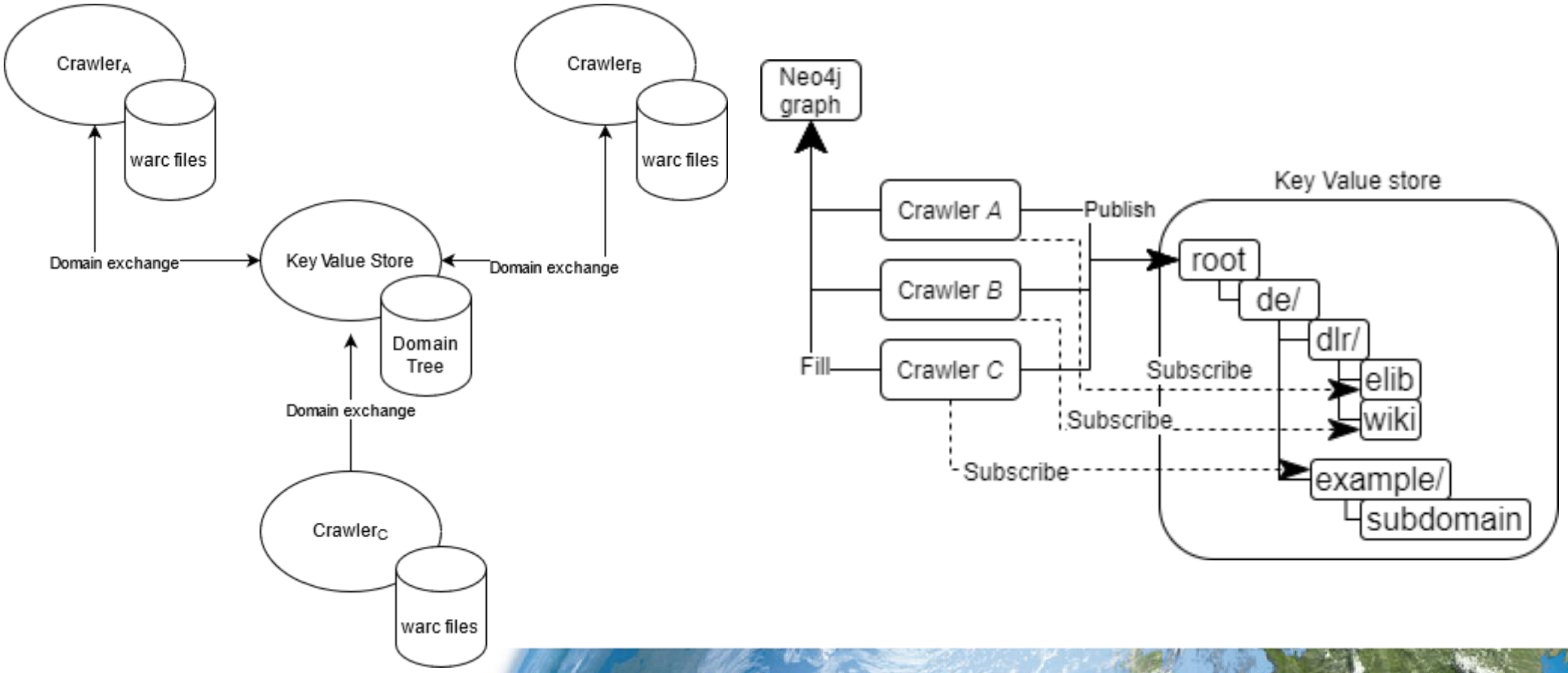
4. Open Search Concept paper for DLR

- 4.1. Concept paper

DLR

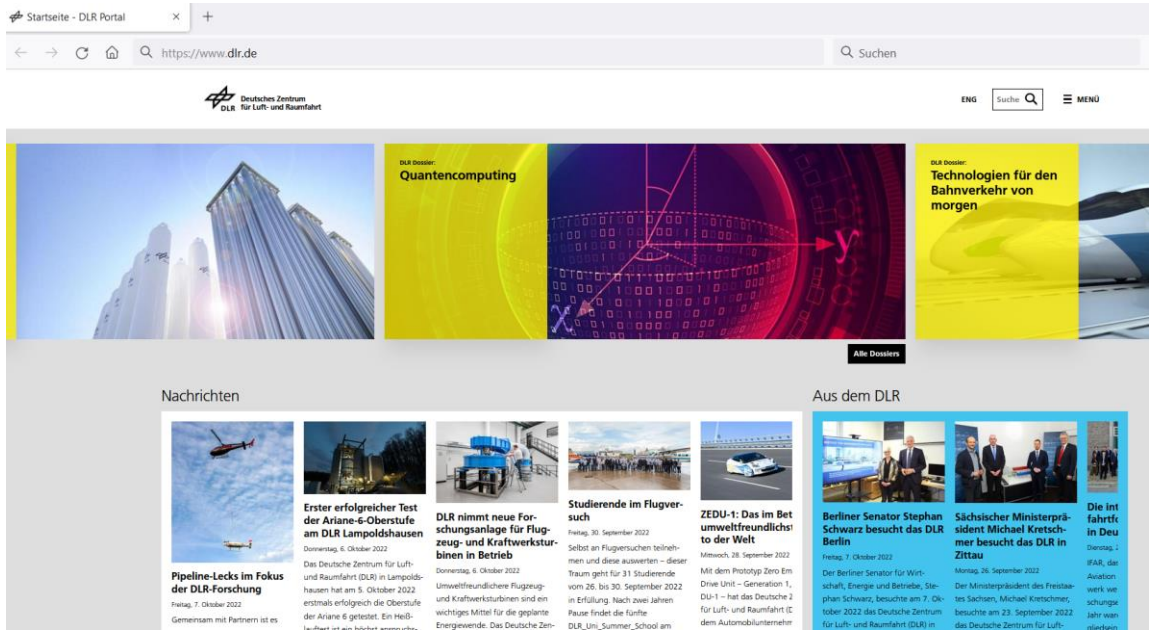
Distributed Crawler Architecture

- Details: https://gitlab.com/opensearch-dlr/opensearch-prototype/-/wikis/home

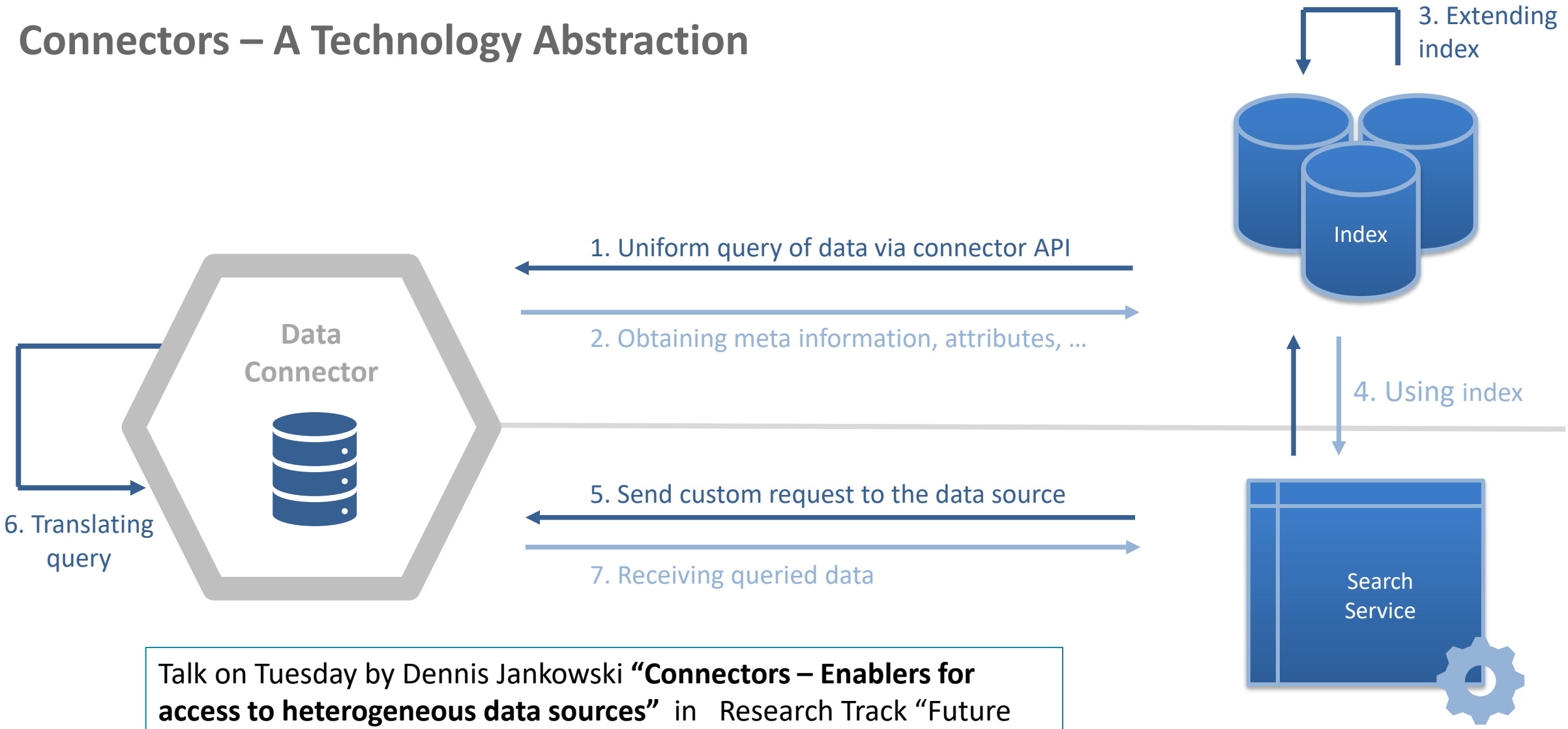


Crawl Results for Subdomain *.dlr.de*

- First crawl: 2021-10-04 08:19:49+00:00
- Latest crawl: 2022-08-26 10:39:07+00:00
- Overall crawled sites: 1.349.830 (4127 sites/day on average)
- Unique crawled sites: 340.501
- Every crawled site is re-visited every 24h, except other policies given in robots.txt.
- Coordination via Elastic Index
  - Every crawler stores which pages were retrieved and when. Other crawlers check for tasks.



Connectors – A Technology Abstraction

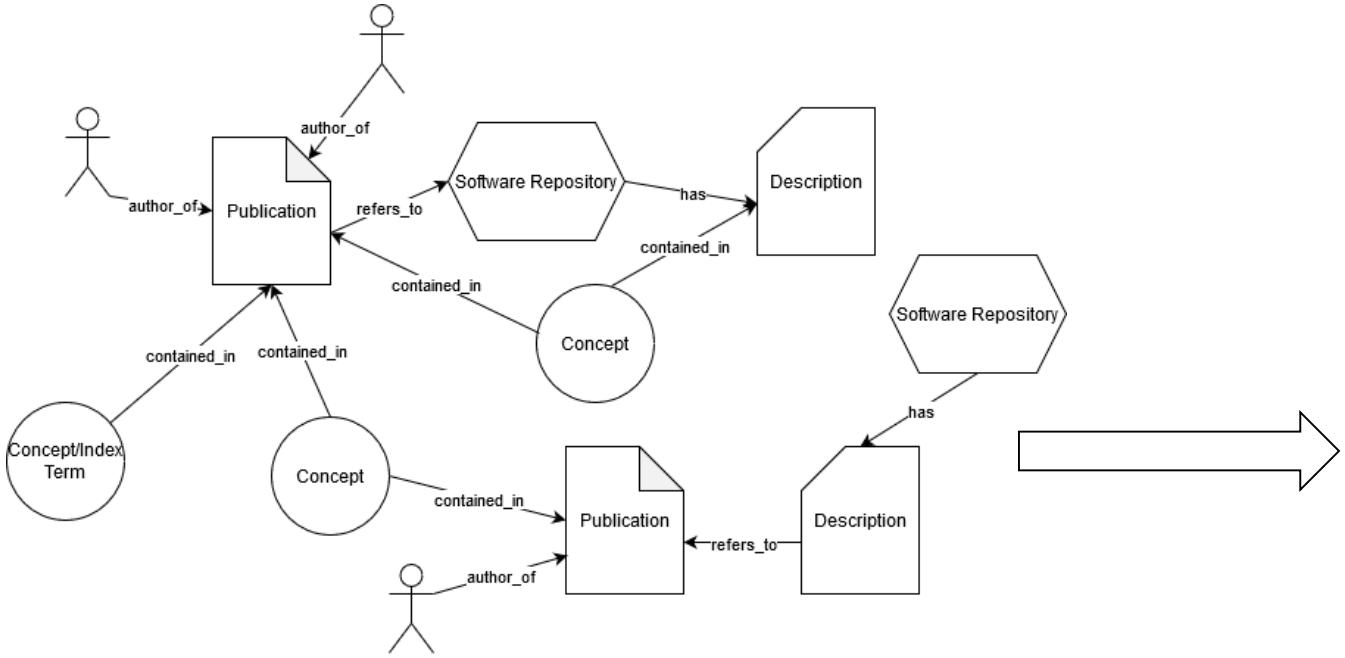


Talk on Tuesday by Dennis Jankowski **“Connectors – Enablers for access to heterogeneous data sources”** in Research Track “Future open search paradigms and concepts” from 15:45

DLR

DLR Open Science Search Prototype

http://app.opensearch.sc.dlr.de/ (DLR internal)



Opensearch @ DLR

sipMask

Publications | Git Repositories | Expert Search | Co-Author Recommendation

Related DLR-Sites

Publication entry

SipMask: Spatial Information Preservation for Fast Image and Video Instance Segmentation
2020

Authors:

Name	Email
Jiale Cao Suggest Co-Authors	-
Rao Muhammad Suggest Co-Authors	-
Hisham Cholakkal Suggest Co-Authors	-
Fahad Shahbaz Suggest Co-Authors	-
Yanwei Pang Suggest Co-Authors	-
Ling Shao Suggest Co-Authors	-

Expert search and collaboration recommendation

Associated Git-projects:

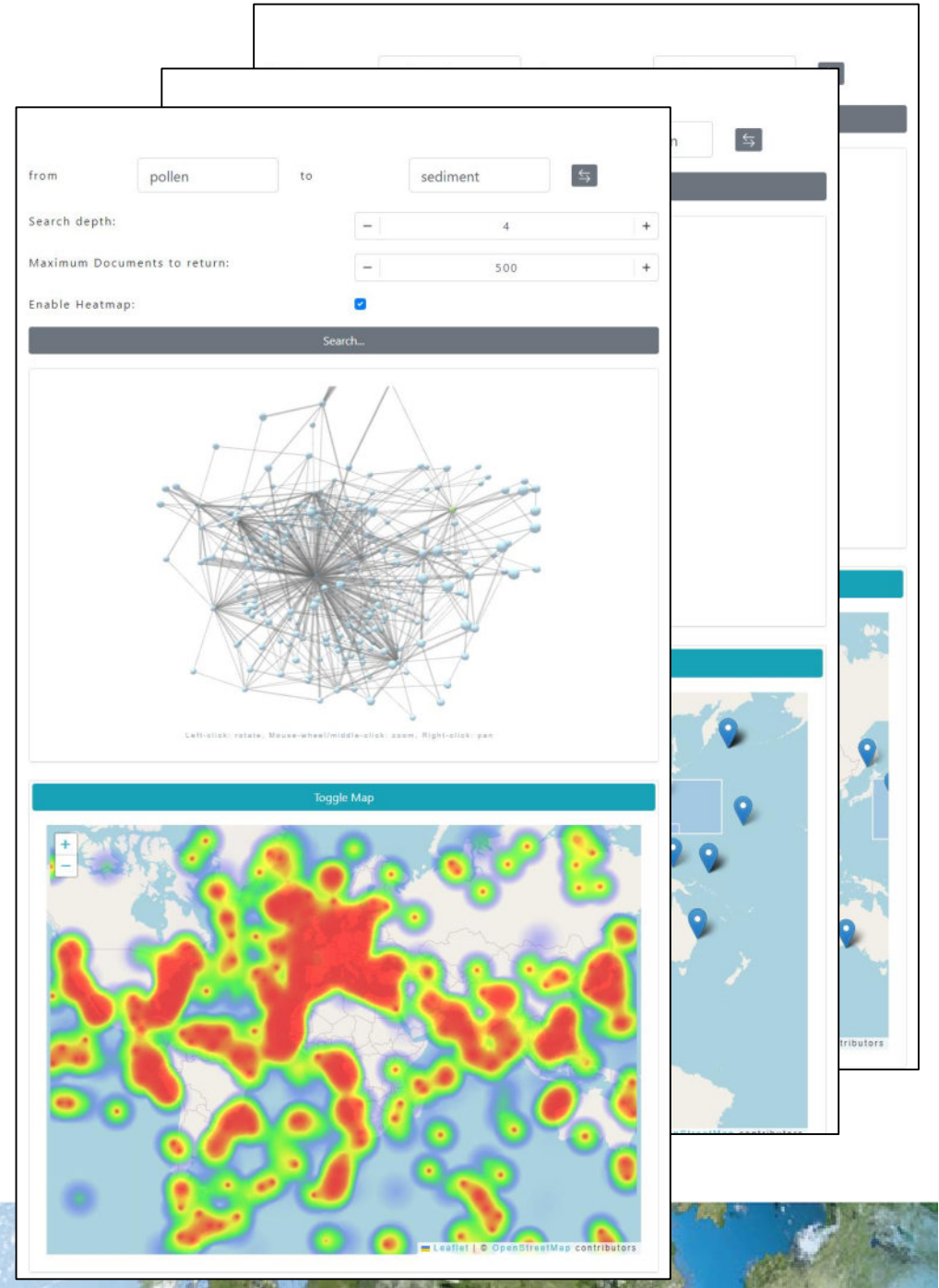
SipMask

Links to related software and datasets

DLR Open Science Search Prototype

http://app.opensearch.sc.dlr.de/ (DLR internal)

- **Dataset search and discovery**
  - Pangaea environmental science dataset and study search
  - Discovery of possible indirect connections between scientific concepts (Literature-based discovery)
  - Location of relevant environmental studies along with references to open datasets



OpenSearch@DLR Workshops, Colloquia, Sprints, Networking…

Past colloquia in 2021 - overview

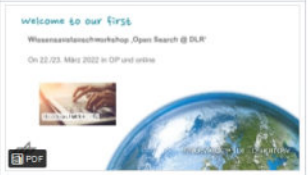
Date/Time	Topic/Speaker
13.04.2021 / 14:00-14:45	"OpenSearch@DLR – Hintergrund und Ziele des Projekts" - Dr. Stefan Voigt, DLR
08.06.2021 / 14:00-14:45	"Practical Experiences and New Challenges in Web Crawling" Christopher Schröder and Martin Potthast - Universität Leipzig / Webis Group
14.09.2021 / 14:00-14:45	"DLR OpenSearch Testbed and first Experiments" - OpenSearch@DLR Team
09.11.2021 / 15:00-15:45	"Semantic representation and analytics for geospatial data" - Prof. Dr. Elena Demidova - Computer Science Institute - University of Bonn

Dates in 2022

Date/Time	Topic/Speaker
15.02.2022 / 14:00-14:45	"Information Extraction and Entity Linking for Semantic Web Search" - Dr. Faegheh Hasibi - Institute of Computing and Information Sciences - Radboud University, NL
05.04.2022 / 14:00-14:45	"The Europe Media Monitor: web news as unconventional data sources" Eng. Marco Verile - Team Leader of the Europe Media Monitor (EMM) at the European Commission – Joint Research Centre (JRC)
24.05.2022 / 14:00-14:45	"A Super Computing Centre Perspective on Open Search" Prof. Dr. Dieter Kranzlmüller, Leibniz Supercomputing Centre (LRZ)
19.07.2022 / 14:00-14:45	"Open Search Pilot Applications at DLR" Dr. Tobias Hecking and Dennis Jankowski (DLR)
27.09.2022 / 14:00-14:45	"Science Search / Open Search @ CERN" Dr. Andreas Wagner, CERN
08.11.2022 / 14:00-14:45	**"Economic Dimension of Open Search" Dr. Oliver Blanchard (Open Search Foundation) and Klaus Fuest (Roland Berger)**
06.12.2022 / 14:00-14:45	"Two years of OpenSearch@DLR – looking backward and ahead" Dr. Stefan Voigt (DLR)

WAW Open Search @ DLR

Day 1: 22 Mar. 2022 (Tuesday)

Time	Agenda item	Description	Presenter	Online / OP	Slides
12:45	Checkin - Tech-Setup		all		
13:00 - 13:45	Opening Session	Welcome, organisational matters and introduction to the Open Search Initiative and Project	Voigt, Stefan	Online/OP	
13:45 - 14:00	DLR Opening Statement	DLR perspective on Open Search	DLR Vorstand Prof. Karsten Lemmer	Online	
14:00-14:45	Invited Keynote Speech	The European Dimension of Open Search	Prof. Michael Granitzer, University Passau	Online	
14:45 - 16:45	Market Place / Poster Session	Getting to know each other and networking	Haupt, Carina	Online (Gather.Town)	
16:45 - 17:45	Scientific Programme - *Geospatial search and web data analysis*	**Registered talks** Scientific talks (10-15 min.) / Lightning talks (~5 min.). See Call for Talks	Hecking, Tobias	Online/OP	

Research-Proposals

Open WebSearch .eu

Research
Information Retrieval, AI, HCI, Geo-spatial Data Processing

UNIVERSITÄT PASSAU
Webis.de
Bauhaus-Universität Weimar
Radboud Universiteit
DLR
TU Graz

Infrastrcuture Organisations
Data Storage, HPC, Services and Scientific Computing

lrz
CERN
VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER
CSC ICT Solutions for Brilliant Minds

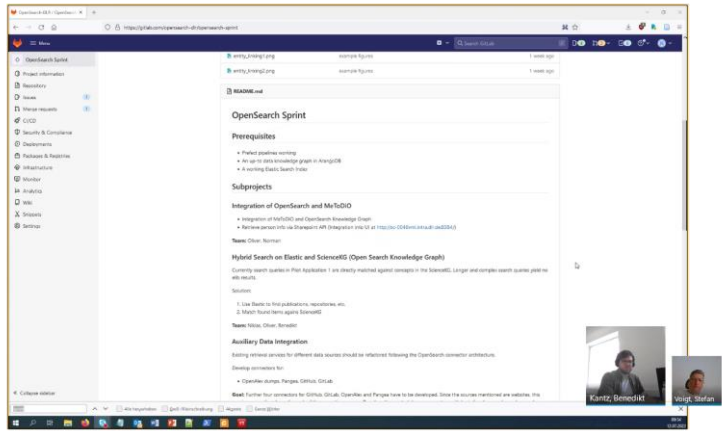
Associations
for a future Web /Internet

open search foundation
nlnet FOUNDATION

Companies
for a future Web /Internet

A1 metaGer

OpenSearch@DLR Coding Sprint



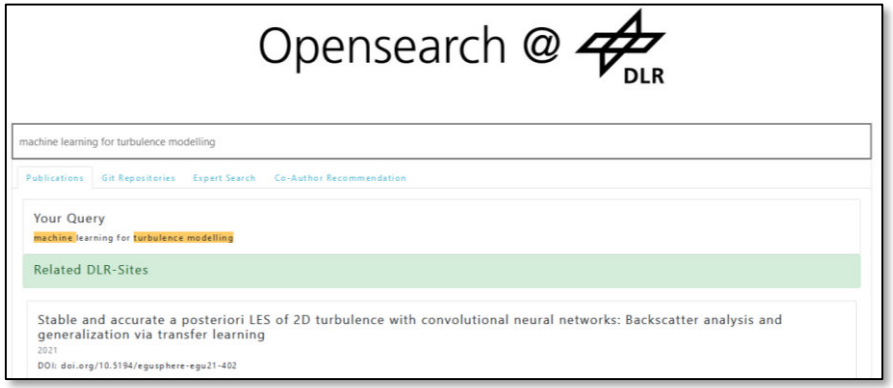
During the OpenSearch@DLR project we found the following (1/2):

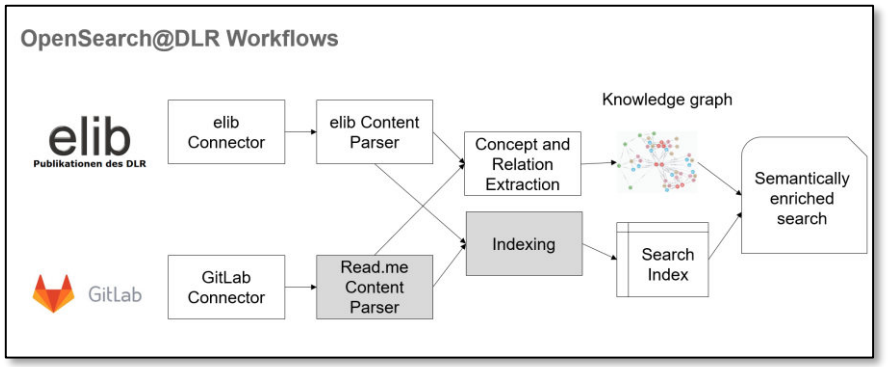
- Efficient **searching and finding of information within the different repositories** within the organization is as important as searching for external information and data on the web.

- In particular for scientists**,** **finding and retrieval of research data** is very important and needs better technical tools and support. This **requires thematically pre-structured and pre-processed information** on existing data and information to better search in internal repositories and corpora.

- Often enough**,** **information needs to be found in heterogeneous artefacts** (such as papers, software tools, manuscripts, etc.) which are often distributed over various different subsystems.

- **Techniques for automatic content analysis** have to be advanced to identify **relationships between scattered pieces** of information and data, which enable integrated **search and discovery of scientific knowledge simultaneously** and synergistically across different repositories and corpora: e.g. to search for authors, scientific concepts, tools or techniques.





DLR

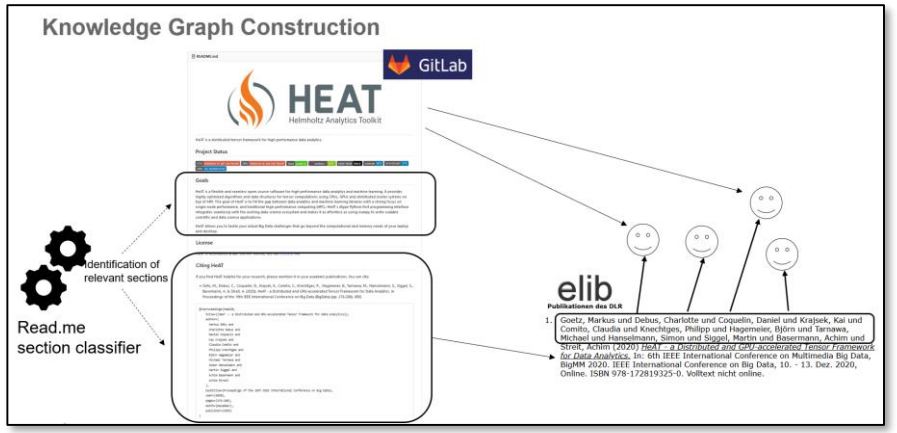
During the OpenSearch@DLR project we found the following (2/2):

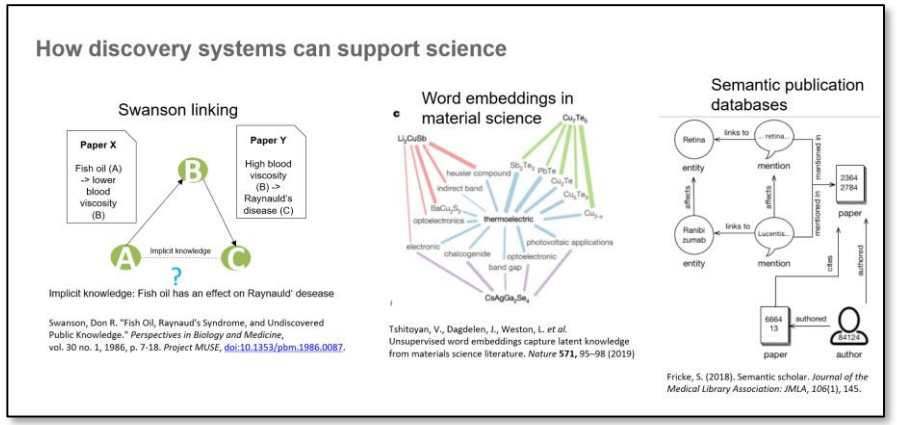
- Sophisticated search involves **efficient visualization and analysis tools**, allowing deeper insight in to search results and going far beyond of what is offered in today's standard search tools.

- This includes for example that the availability and analysis capabilities of **geospatial search features need to be significantly improved** for all kinds of search and analysis tasks in the web. Along with this, time-stamps and versioning of web artefacts are equally important factors.

- Versions of knowledge artefacts have to be tracked in order to avoid inconsistencies and ensure that data of **known provenance and up-to-date-ness are used for scientific analysis.**

- It is **desirable to identify emerging topics and to anticipate future trends of scientific and technological or even societal developments.** Representing extracted information and resources as well as **semantic relations** between them in **graph-databases** seems to be a promising approach for this.





In conclusion:

- All in all, it can be established that **good internal search features**, good internal **knowledge management** and thus, **sophisticated intranet search** is a critical infrastructure for science centers, which need better care and which by far should not be considered a by-product.

- In conclusion we argue that improving the **intranet search features and linking them with sound public and distributed open web search** infrastructures will substantially improve the scientific work and organizational efficiency of many large-scale science organizations.

- Thus, any major **science and research organization** needs to actively engage in the building and maintaining of **openly searchable information repositories internally**, as well as where possible, **also externally**, in cooperation with others and for searching the web as a whole.

contact: opensearch@dlr.de