

Duplicate Evaluation in the European Data Portal

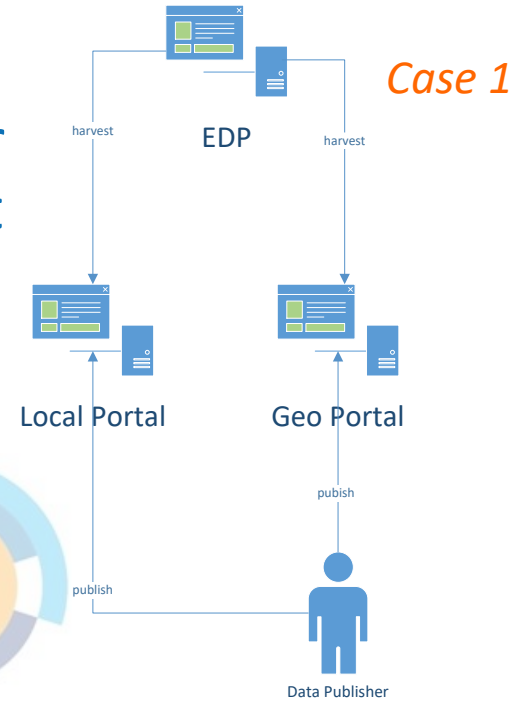


EUROPEAN
DATA PORTAL

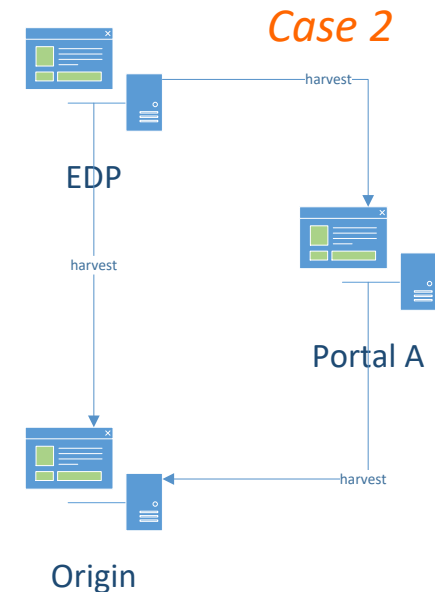
Simon Dutkowski – Fraunhofer FOKUS
SDSVoc 2016



- Case 1: Simply because the provider publishes a dataset on two different portals



- Case 2: Harvesting the originating portal and another portal that harvests the originating portal too.



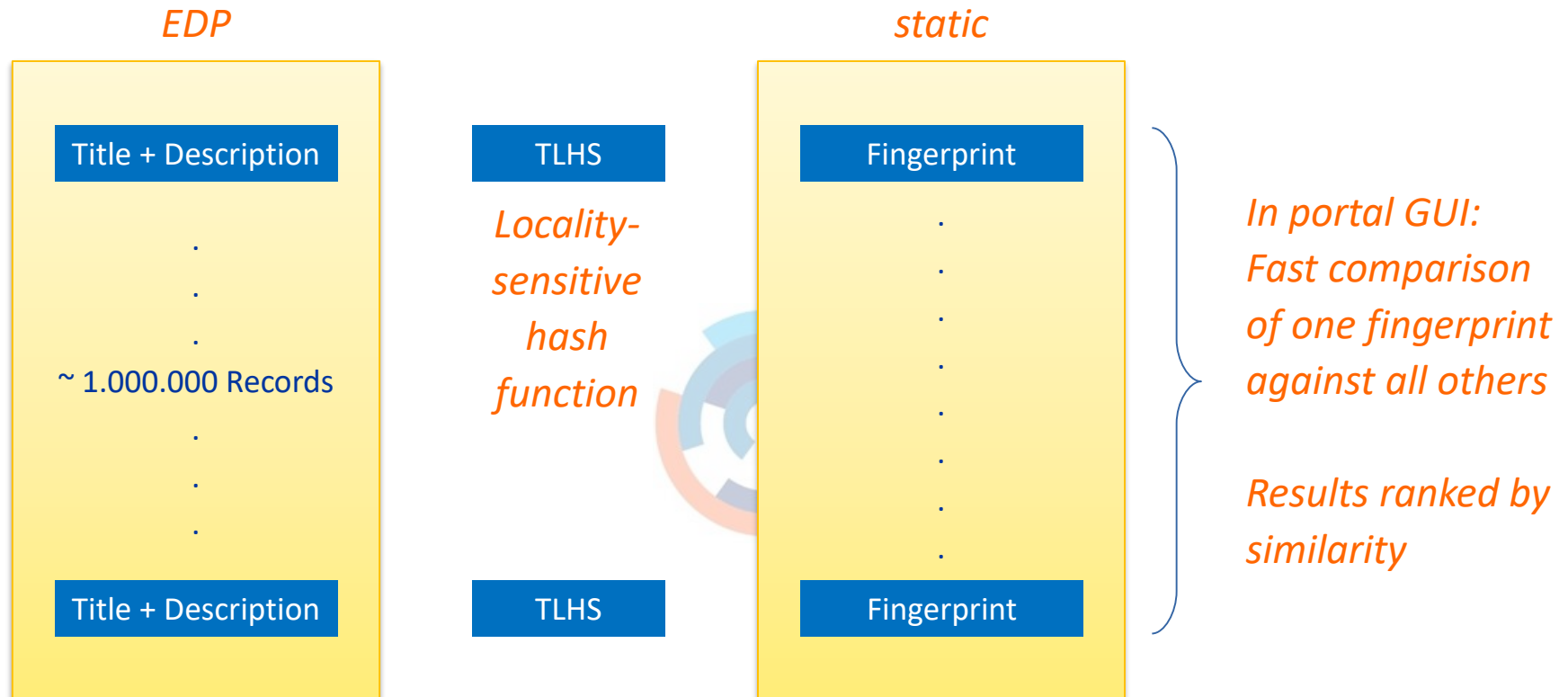
Reasons for detecting and eliminating duplicates

- 🤖 First of all for better user experience. Search results should not contain duplicates.
- 🤖 Avoiding unnecessary resource consumption. In case 2 we probably have a complete catalogue duplicated
- 🤖 No 100% safe method to detect duplicates as long as there is more than one metadata standard involved.
- 🤖 During transformation identifiers or provenance information can be lost.



- 🕒 Detection of duplicates, by text similarity
- 🕒 Fast pairwise comparisons of titles + descriptions
– order of millions –
- 🕒 Vector space model / cosine distance: too slow
- 🕒 locality-sensitive hashing TLSH:
short fingerprints, fast distance function


Fingerprinting and Ranking










- 🐙 Our texts too short for TLSH
 - Successful adaptation of TLSH (#buckets 256 → 64)
- 🐙 TLSH too slow for pairwise comparisons of millions
 - Limitation to compare only within language areas
- 🐙 TLSH fast ($<1\mu\text{s}$ per comparison)
 - Fast enough for comparison of a single text with all others in GUI time scale
- 🐙 Text similarity not sufficient for detecting duplicates
 - 🐙 Almost identical texts, e.g. with the word “gas” in one of them and “electricity” in the other.

EDP Similarity Search Demo

Search Maintenance

renewable  Suchen







find similar datasets (by title+description)

Rank	Dataset Information	Similarity Match
1.	DE ...541a0c7b-134c-0432-cfb8-6cd47314f239 title: Level 2013 — Renewable energy: Stock of Fotovoltaikanlagen, number (circle) ¶ text length: 357 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
2.	DE ...e6f3accb-f85e-6736-36e9-f59954e84606 title: Level 2013 — Renewable energy: Electricity from photovoltaics in kilowatt-hours (municipal) ¶ text length: 393 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
3.	DE ...9206fd4b-5df8-605f-89ca-bcff625b04e8 title: Level 2013 — Renewable energy: In kilowatt-hours of electricity from wind power verbandsgemeindeebene) ¶ text length: 409 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
4.	DE ...41eb53ce-d132-07f8-825c-8870b3d78462 title: Level 2013 — Renewable energy: In kilowatt-hours of electricity from renewable energy sources (circle) ¶ text length: 421 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
5.	DE ...005bb0f4-1e96-c921-a023-e581be8c3ba8 title: Level 2013 — Renewable energy: In kilowatt-hours of electricity from biomass (circle) ¶ text length: 387 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
6.	DE ...ldbnrw-service-1728438067ldb title: NRW: Completed construction projects (new building and alterations to existing buildings): Residential and non-residential buildings, flats, rooms, accommodation and land — municipalities — year ¶ text length: 590 publisher: Information und Technik Nordrhein-Westfalen update: ??x	
7.	DE ...ldbnrw-service-1728437106ldb title: NRW: Completed construction projects (new building and alterations to existing buildings): Residential and non-residential buildings, dwellings by number of rooms, areas, municipalities — year ¶ text length: 594 publisher: Information und Technik Nordrhein-Westfalen update: ??x	

Original and similar dataset: dc:title+description w/o stopwords

Level 2013 — Renewable energy: In kilowatt-hours of electricity from renewable energy sources (circle) ¶
Thematic maps on renewable energy. Demonstrate input from renewables such as photovoltaics, wind power, hydropower and biomass in kilowatt-hours at district level, as well as the number of photovoltaic and wind turbines and their rated power at county and verbandsgemeindeebene. In kilowatt-hours of electricity from renewable energy sources, district level

Level 2013 — Renewable energy: In kilowatt-hours of electricity from biomass (circle) ¶
Thematic maps on renewable energy. Demonstrate input from renewables such as photovoltaics, wind power, hydropower and biomass in kilowatt-hours at district level, as well as the number of photovoltaic and wind turbines and their rated power at county and verbandsgemeindeebene. In kilowatt-hours of electricity from biomass, district level

13	DE ...005bb0f4-1e96-c921-a023-e581be8c3ba8 title: Level 2013 — Renewable energy: In kilowatt-hours of electricity from biomass (circle) ¶ text length: 387 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
13	DE ...444280ac-9249-3f60-4f90-bce0fb3c211b title: Level 2013 — Renewable energy: In kilowatt-hours of electricity from hydropower (circle) ¶ text length: 393 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
16	DE ...8d789140-a518-df00-f4e5-34277230c681 title: Level 2013 — Renewable energy: In kilowatt-hours of electricity from wind power (circle) ¶ text length: 401 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
22	DE ...9206fd4b-5df8-605f-89ca-bcff625b04e8 title: Level 2013 — Renewable energy: In kilowatt-hours of electricity from wind power verbandsgemeindeebene) ¶ text length: 409 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
23	DE ...02e14fca-872c-877d-b5c5-7599cb303f16 title: Level 2013 — Renewable energy: In kilowatt-hours of electricity from PV verbandsgemeindeebene) ¶ text length: 413 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	
23	DE ...0b1a3e7d-7b23-f87b-9cd1-48f2b0e6a543 title: Level 2013 — Renewable energy: In kilowatt-hours of electricity from photovoltaic (PV) ¶ text length: 387 publisher: Statistisches Landesamt Rheinland-Pfalz update: ??x	

... show more

20279 RAM records compared; 85 similarity match(es); in 14 milliseconds

DE



EUROPEAN
DATA PORTAL

Simon Dutkowski

- ▶ Fraunhofer FOKUS
- ▶ simon.dutkowski@fokus.fraunhofer.de
- ▶ +49 30 34637128