



Making Inspire data discoverable and findable through popular search engines

Abdelfettah Feliachi, Sylvain Grellet, Thierry Vilnius

► To cite this version:

Abdelfettah Feliachi, Sylvain Grellet, Thierry Vilnius. Making Inspire data discoverable and findable through popular search engines: The FRENCH Experimentation on geocatalogue : IT context , French context, Proposed data structure, URIs in the picture, Implementation for Datasets, URIS – applied to the DATA structure, Example of indexation in Search engines. INSPIRE Discovery workshop, Jul 2019, ISPRA, Italy. hal-03925452

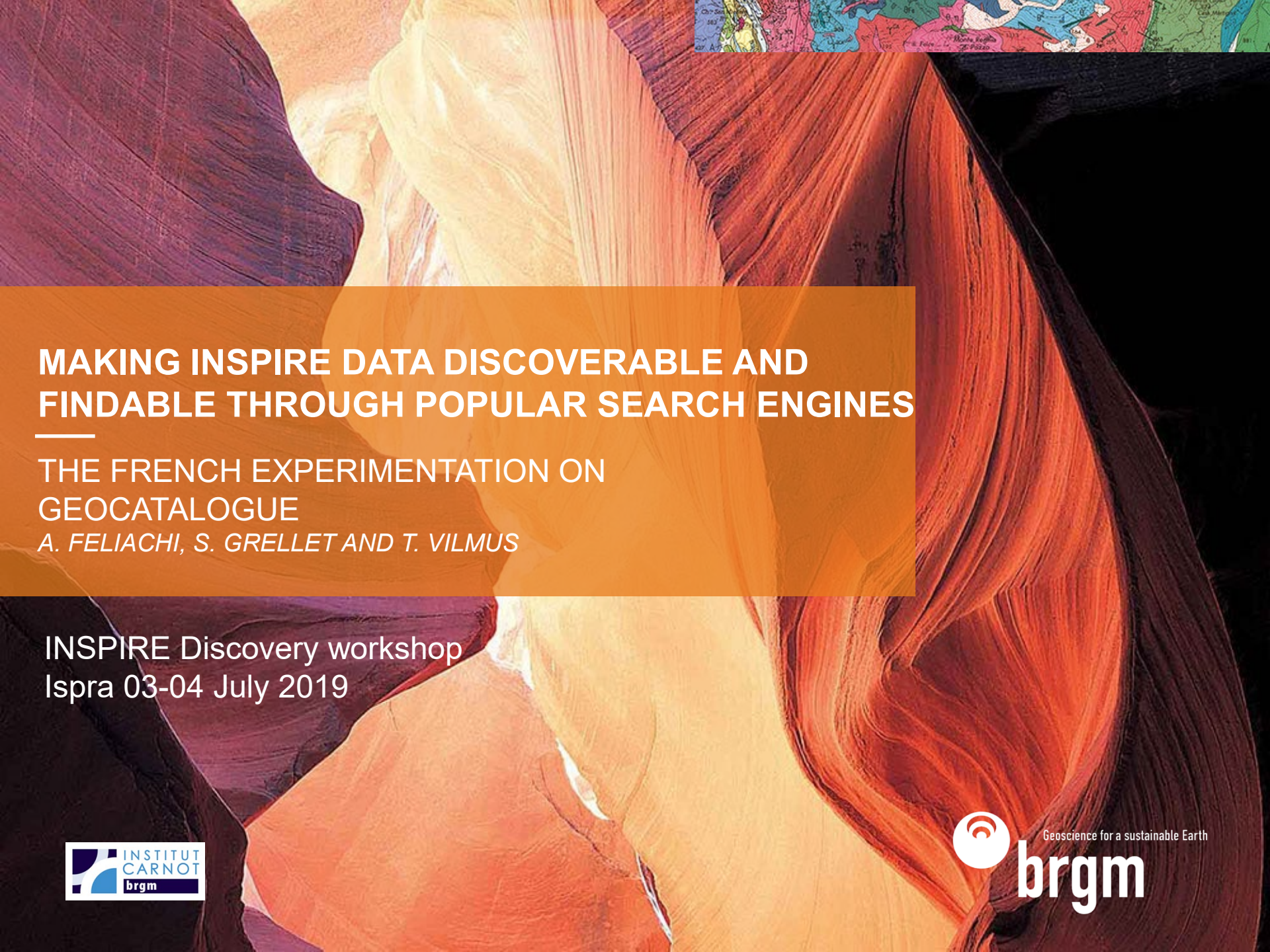
HAL Id: hal-03925452

<https://brgm.hal.science/hal-03925452>

Submitted on 5 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



MAKING INSPIRE DATA DISCOVERABLE AND FINDABLE THROUGH POPULAR SEARCH ENGINES

THE FRENCH EXPERIMENTATION ON
GEOCATALOGUE

A. FELIACHI, S. GRELLET AND T. VILMUS

INSPIRE Discovery workshop
Ispra 03-04 July 2019

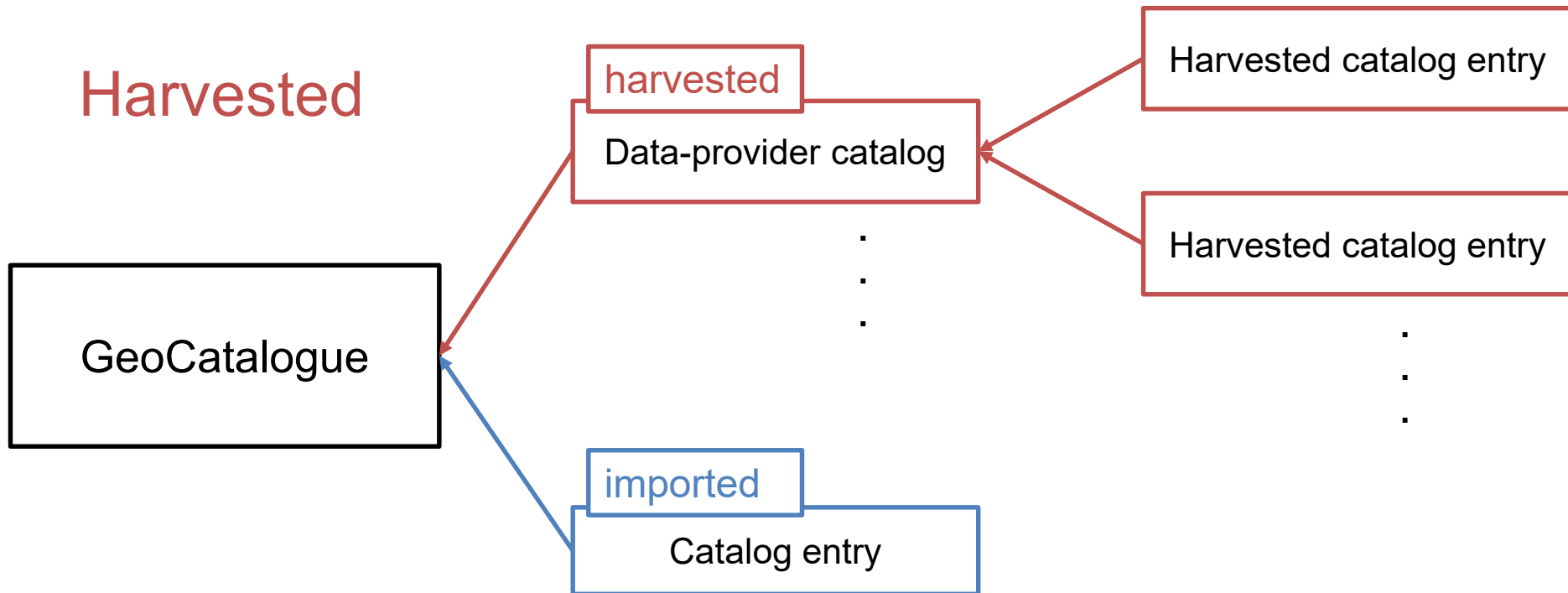
IT CONTEXT

- BRGM, French geological survey, is implementing the national INSPIRE catalogue, named GeoCatalogue
- It's hard to find datasets
 - Difficulties to find data through Inspire specialized search engines like Geoportals or Geocatalogs
 - General public even unaware of the existence of such tools
- How to help search engine index those datasets ?
 - Vocabulary : Schema.org
 - Proposed by important search engines Google, Microsoft, Yahoo and Yandex
 - Payload
 - JSON-LD embedded in HTML pages

FRENCH CONTEXT

- Metadata flow into GeoCatalogue

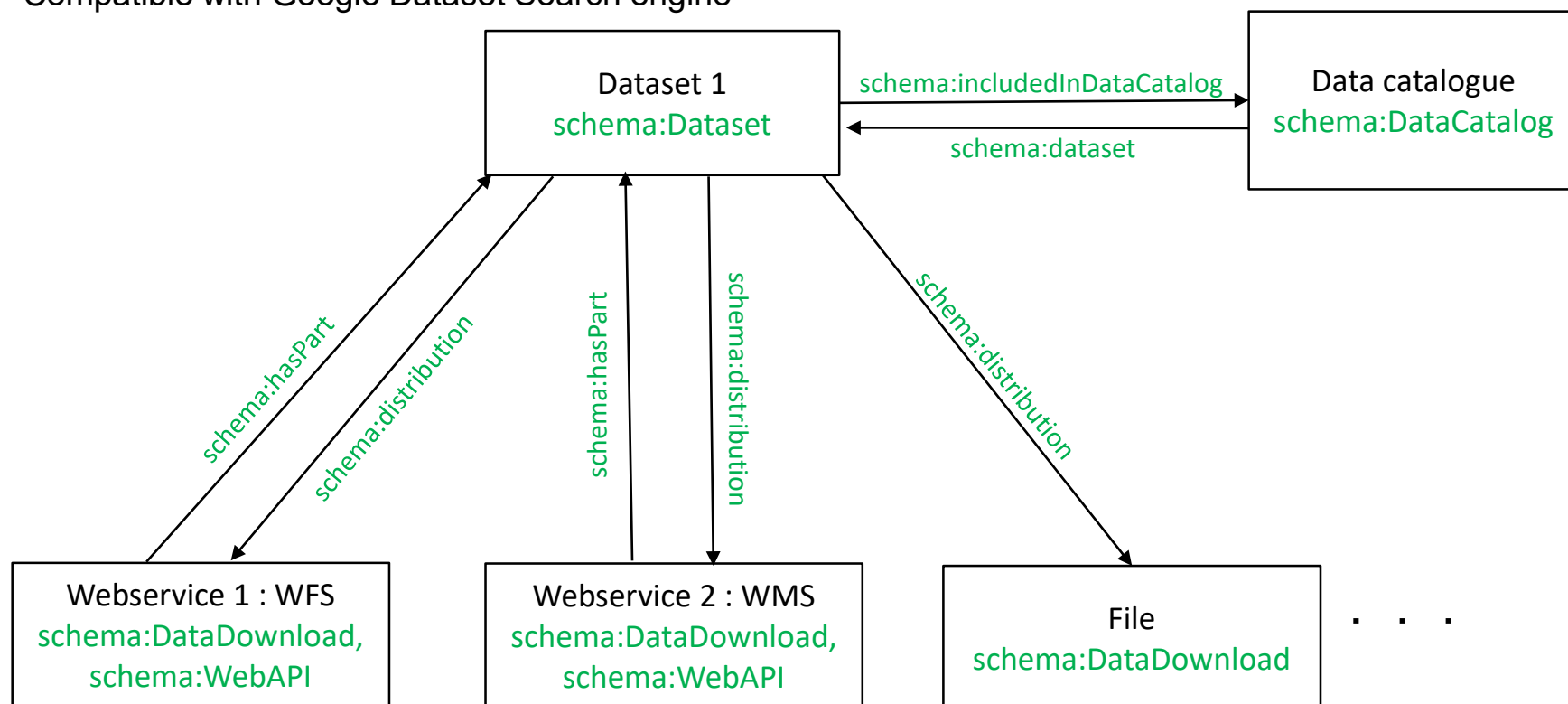
Harvested



Manually imported

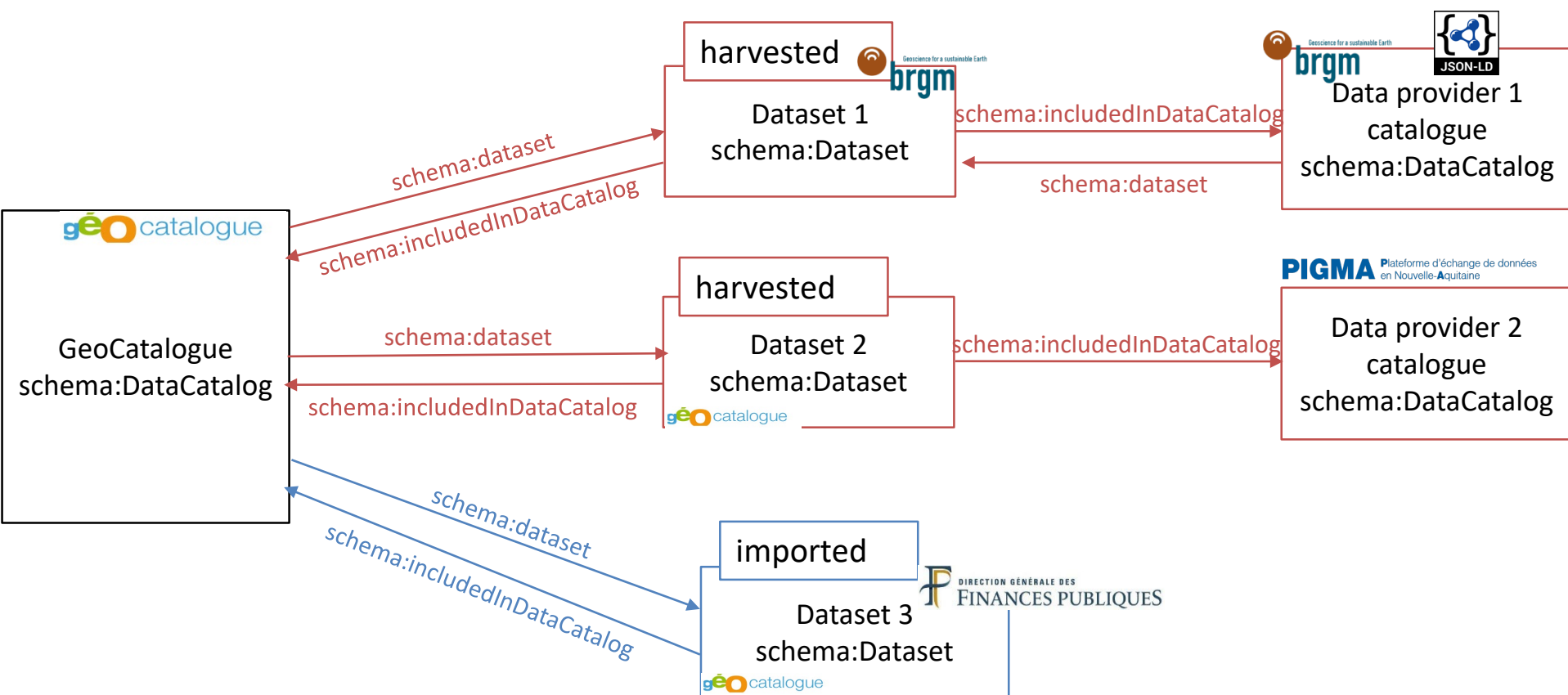
PROPOSED DATA STRUCTURE

- Generic JSON-LD approach : as recommended by search engines
- Use schema:dataset & schema:includedInDatacatalog to link catalogues and datasets
- Use schema:distribution to declare services
- Compatible with Google Dataset Search engine



PROPOSED DATA STRUCTURE

- Not all data provider have a URI policy that resolves to a well defined JSON-LD representation
- Example below



PROPOSED DATA STRUCTURE

● JSON-LD examples (dataset, catalogue & service)

Catalogue

```
{
  "@context": "http://schema.org/",
  "@type": "DataCatalog",
  "@id": "https://data.geoscience.fr/id/catalogue/BRGM",
  "name": { "value": "BRGM Data Catalog", "@language": "en" },
  "description": "BRGM metadata catalog",
  ....
  "dataset": [ "https://data.geoscience.fr/id/dataset/borehole", ... ]
  ....
  "about": [ "https://www.eionet.europa.eu/gemet/en/inspire-theme/ge", ... ],
  ....
}
```

Dataset

```
{
  "@context": "http://schema.org/",
  "@type": "Dataset",
  "@id": "https://data.geoscience.fr/id/dataset/borehole",
  "includedInDataCatalog": "https://data.geoscience.fr/id/catalogue/BRGM",
  "name": { "value": "Borehole", "@language": "en" },
  ...
  "distribution": [
    { "@id": "https://data.geoscience.fr/api/wfs/borehole",
      "@type": [ "DataDownload", "WebAPI" ],
      "contentUrl": "http://geoservices.brgm.fr" } ... ],
  ....
}
```

Service

```
{
  "@context": "http://schema.org/",
  "@id": "https://data.geoscience.fr/api/wfs/borehole",
  "@type": [ "DataDownload", "WebAPI" ],
  "name": "Borehole WFS Service",
  ....
  "keywords": [
    { "@value": "Forage", "@language": "fr" }, ... ],
  ....
  "spatialCoverage": { "@type": "Place",
    "geo": { "@type": "GeoShape",
      "box": [ "-5.79028,41.36493 9.56222,51.09111",
        "-61.7961,15.87 -61.1871,16.5129",
        "-61.2315,14.4028 -60.817,14.8801",
        "-54.6038,2.11347 -51.6481,5.75542", *
        "55.2206,-21.3739 55.8531,-20.8565",
        "45.0392,-12.9925 45.2297,-12.6625" ]
      }
    }, .....
}
```

URIS IN THE PICTURE

- Define a national URI architecture
 - Taking into account the 3 types of data providers
 - Harvested by the national catalogue : with a URI policy & with no URI policy
 - Imported into the national catalogue (thus no data provider URI policy)
- Use persistent URI to identify catalogues, datasets and services
- Rationale
 - For data provider having a URI policy that resolves in JSON-LD : respect it
 - For the others : define a national pattern

Data catalogue : https://data.geocatalogue.fr/id/catalog/{data_provider_catalogue_id}

Dataset : https://data.geocatalogue.fr/id/dataset/{geocatalogue_defined_uuid}

Handled through a unique URI resolver

→ When those start having a URI policy that resolves in JSON-LD have a HTTP 301 ('Moved Permanently') from the previous URI to the new one

IMPLEMENTATION FOR DATASETS

- ISO 19115 (19139 XML encoding) to JSON-LD/Schema.org mapping
 - Building on feedback from previous experience:
https://www.w3.org/2015/spatial/wiki/ISO_19115_-_DCAT_-_Schema.org_mapping
<https://ec-jrc.github.io/dcat-ap-to-schema-org/>
<http://geocat.fr/dataset-prop.html>
 - Proposal of an operational mapping
<https://github.com/geonetwork/core-geonetwork/wiki/JSON-LD---ISO19139-mapping-proposal>
- XSLT implementation experimentation
 - On the fly generated JSON-LD from the 19139 XML encoding of the metadata
 - Imbedded in the HTLM pages

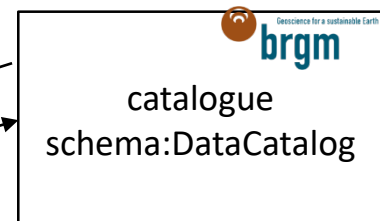
URIS – APPLIED TO THE DATA STRUCTURE

- Data provider with a URI policy that resolves to JSON-LD
- Comprehensive example on BRGM national borehole dataset

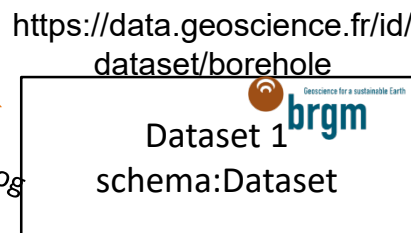
<https://data.geocatalogue.fr/id/catalog/geocatalogue>



<https://data.geoscience.fr/id/catalogue/BRGM>

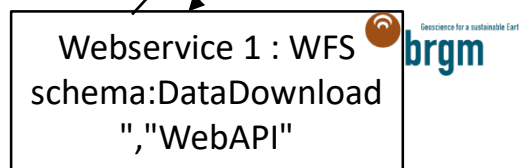


schema:dataset
schema:includedInDataCatalog



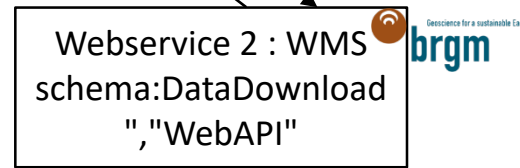
schema:dataset
schema:includedInDataCatalog

schema:hasPart
schema:distribution



<https://data.geoscience.fr/api/wfs/borehole>

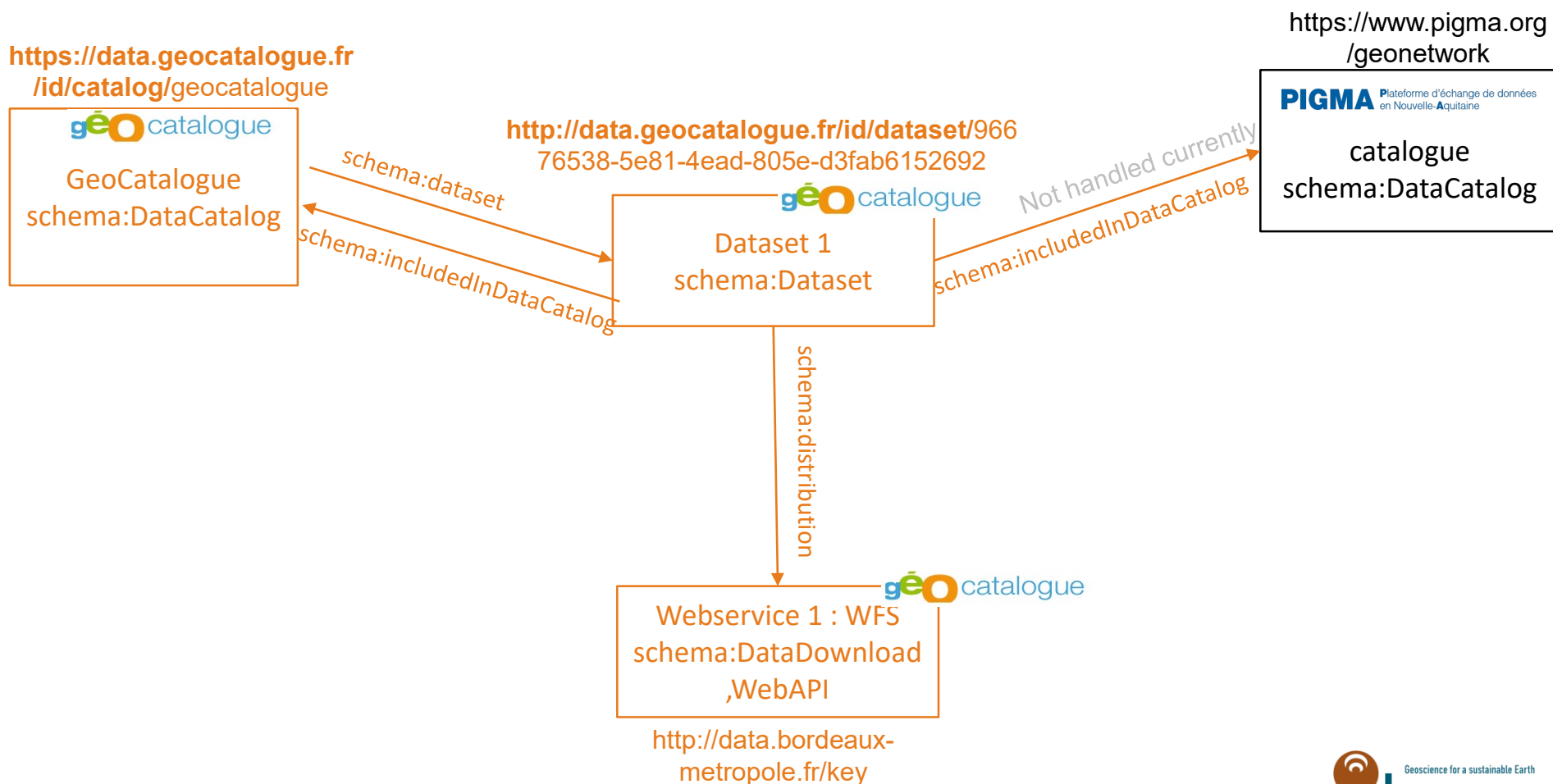
schema:distribution
schema:hasPart



<https://data.geoscience.fr/api/wms/borehole>

URIS – APPLIED TO THE DATA STRUCTURE

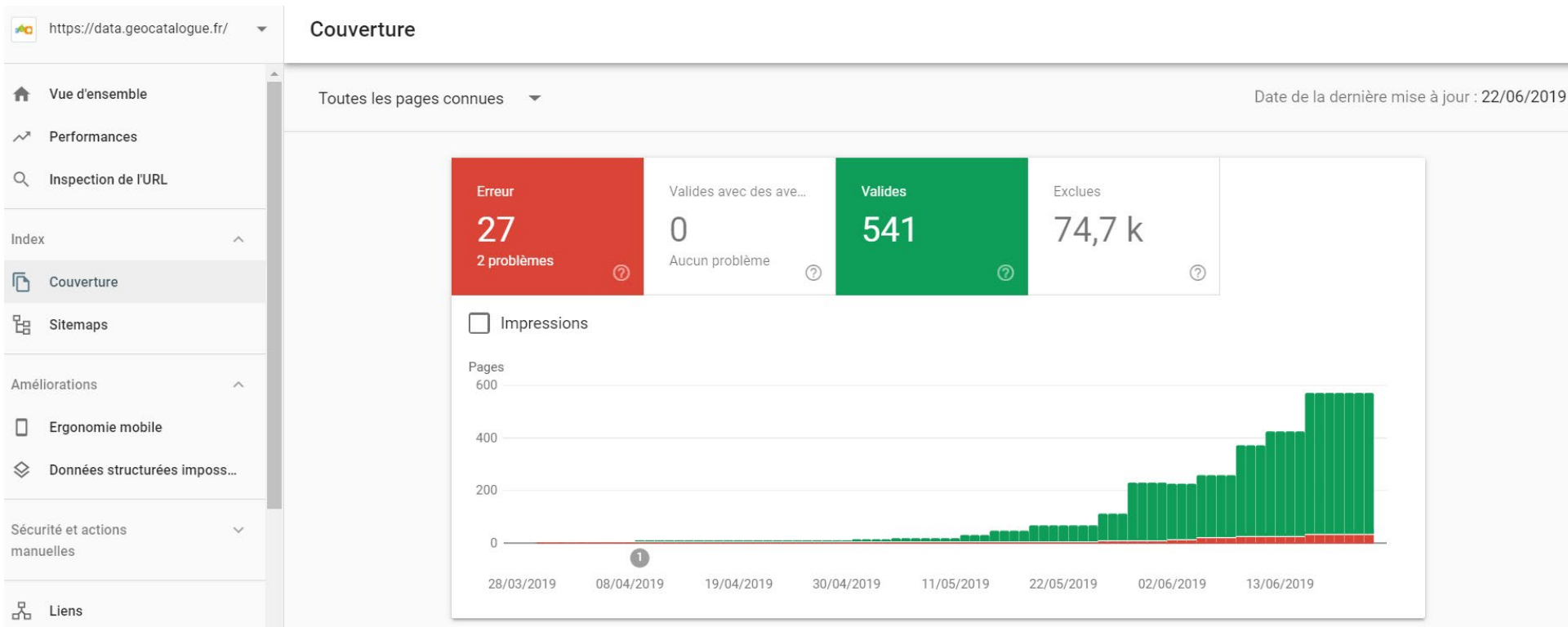
- Data provider with non URI or a URI policy that does not resolve to JSON-LD
- Comprehensive example on PIGMA platform



EXAMPLE OF INDEXATION IN SEARCH ENGINES



- Google search console
 - Sitemap needed for indexation: generating file(s) manually periodically
 - Uploading them to the Google search console
 - Run indexation then wait ...



WHAT'S NEXT

- Pending IT aspects
 - How to declare a webservice that is not linked to a specific dataset (ex : WPS) ?
 - Link from catalogue to catalogue ?
 - Follow DCAT2 / schema.org work
 - Possibility to use vocabulary from dcat (ex: dcat:DataService, ...) : how is it indexed by search engines

- Implementation
 - Basic SEO must be respected. (HTML title corrected recently to correspond the dataset title)
 - Improve the XSLT JSON-LD generation: some errors are detected by the search engine console.
 - Agree on JSON-LD patterns for services
 - Follow / Finish the test of the architecture at national scale
 - Push the solution to open source projects (ex : Geonetwork)

CONCLUSION

- Indexation results
 - Google Search : Slow but effective. Pages do not necessarily hit the first page. Adding the term “geocatalogue” to the search improves the results.
 - Google Dataset Search : promising results. Searching by dataset name works for the indexed datasets. To explore further: search by keywords, temporal extent, spatial extent, etc.
 - Other search engines: current tests on Bing
- Benefits
 - National GeoCatalogue and linked catalogues : increases usability and visibility
 - Public : enhances overall search experience, allowing to discover, browse, view and download much more environmental data than before.

An important complementary access point for geocatalogs search engines

THANK YOU

Contact

- A.Feliachi@brgm.fr
- S.Grellet@brgm.fr
- T.Vilmus@brgm.fr

