fct arquivo.pt

Inlink dataset walkthrough

FCCN serviços digitais fct

fct Fundação para a Ciência e a Tecnologia

Inlink dataset overview

The Arquivo.pt link dataset combines three distinct web collections:

- **PWA9609** (1996-2009): 89 million pages that capture the initial evolution of the Internet, centered on the .pt domain. This historical collection provides insights into early linking patterns on the Web.
- **AWP38** (Oct-Nov 2021): 44 million pages that offer a contemporary portrait of Web connectivity, with emphasis on the .pt domain, but including broader Internet content.
- **FAWP47** (Oct-Dec 2021): 8 million pages from daily captures of .pt domain content, designed to track short-term changes in linking patterns.

Format description

- **url**: URL (SURT) of the page being linked (linked page)
- **count**: Total number of links to linked page
- **countInternal**: Total number of links to linked page from the same Fully Qualified Domain Name (FQDN)
- **countExternal:** Total number of links to linked page from different FQDN
- **captureDate**: Date time when the linked page was captured in Arquivo.PT (may be null)
- **inlinks**: List of pages that link (linking pages) to the linked page with metadata

Inlinks Object Structure

- **date**: Date time when the linking page was captured in Arquivo.PT
- **source**: URL (SURT) of the linking page
- **anchor**: anchor text for the link in the linking page

Example object

```
{
    "url": "(pt,arquivo,",
    "count": 4,
    "countInternal": 0,
    "countExternal": 4,
    "captureDate": "2021-10-06T20:46:44",
    "inlinks": [
        {
            "date": "2021-12-13T22:25:55",
            "source": "(pt,sapo,blogs,porabrantes,",
            "anchor": "Arquivo da Web portuguesa"
        }, {
            "date": "2021-12-22T19:52:12",
            "source":
"(eu,europa,data,)/sites/default/files/landscaping_insight_report_n7_2021.pdf",
            "anchor": "https://arquivo.pt/"
        }, {
            "date": "2021-12-13T22:25:30",
            "source": "(pt,sapo,blogs,porabrantes,)/o-rosto-de-umbelina-inacio-4275873",
            "anchor": "Arquivo da Web portuguesa"
        }, {
            "date": "2021-10-06T20:27:45",
            "source": "(com,useragentstring,",
            "anchor": "arquivo.pt"
        }
    ]
}
```

How was this created?

- Each webpage has a set of links to other pages, which we’ll call *outlinks*
  - They tell us where a specific page *points* to
- If we are able to reverse this information, we can find which pages *point* to a specific page, and thus, build a link graph that tells us which pages are most often pointed to
- This shares some similarities to PageRank, but only running once

Turning outlinks into inlinks

Extracting pages and inlinks

Doc (url: “https://www.fct.pt”)

- Captured: “2024-03-01 09:00:00”
- Outlinks:
  - (pt.fct/,”Home”)

Processing a WARC, found **fct.pt** page record**,** captured at **2024-03-01 09:00:00**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
    - Captured: “2024-03-01 09:00:00”
    - Outlinks:
        - (pt.fct/,”Home”)

Create new entry in the output map and write doc to output map entry at SURT **pt,fct)/**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)

Found **outlink**, process it into an **inlink**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”

Write **inlink** to **pt,fct)/** entry

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”

- Doc (url: “https://fccn.pt”)
  - Captured: “2024-03-01 10:00:00”
  - Outlinks:
    - (pt.fct/,”Fundação Ciência Tec.”)
    - (pt.fccn/quem-somos,”Quem somos”)

Processing a WARC, found **fccn.pt** page record, captured at **2024-03-01 10:00:00**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”

pt,fccn)/

- Doc #1 (url: “https://fccn.pt”)
  - Captured: “2024-03-01 10:00:00”
  - Outlinks:
    - (pt.fct/,”Fundação Ciência Tec.”)
    - (pt.fccn/quem-somos,”Quem somos”)

Create new entry in the output map and write doc to output map entry at SURT **pt,fccn)/**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink #2
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”

pt,fccn)/

- Doc #1 (url: “https://fccn.pt”)
  - Captured: “2024-03-01 10:00:00”
  - Outlinks:
    - (pt.fct/,”Fundação Ciência Tec.”)
    - (pt.fccn/quem-somos,”Quem somos”)

Found **outlink**, process it into an **inlink** and write to **pt,fct)/**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink #2
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”

pt,fccn)/

- Doc #1 (url: “https://fccn.pt”)
  - Captured: “2024-03-01 10:00:00”
  - Outlinks:
    - (pt.fct/,”Fundação Ciência Tec.”)
    - (pt.fccn/quem-somos,”Quem somos”)

Found **outlink**, process it into an **inlink**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink #2
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”

pt,fccn)/

- Doc #1 (url: “https://fccn.pt”)
  - Captured: “2024-03-01 10:00:00”
  - Outlinks:
    - (pt.fct/,”Fundação Ciência Tec.”)
    - (pt.fccn/quem-somos,”Quem somos”)

pt,fccn)/quem-somos

Create entry at **pt,fccn)/quem-somos**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink #2
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”

pt,fccn)/

- Doc #1 (url: “https://fccn.pt”)
  - Captured: “2024-03-01 10:00:00”
  - Outlinks:
    - (pt.fct/,”Fundação Ciência Tec.”)
    - (pt.fccn/quem-somos,”Quem somos”)

pt,fccn)/quem-somos

- Inlink #1
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Quem somos”

Write **inlink** to **pt,fccn)/quem-somos**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
    - Captured: “2024-03-01 09:00:00”
    - Outlinks:
        - (pt.fct/,”Home”)
- Inlink #1
    - Captured: “2024-03-01 09:00:00”
    - Source: pt,fct)/
    - Anchor: “Home”
- Inlink #2
    - Captured: “2024-03-01 10:00:00”
    - Source: pt,fccn)/
    - Anchor: “Fundação Ciência Tec.”

pt,fccn)/

- Doc #1 (url: “https://fccn.pt”)
    - Captured: “2024-03-01 10:00:00”
    - Outlinks:
        - (pt.fct/,”Fundação Ciência Tec.”)
        - (pt.fccn/quem-somos,”Quem somos”)

pt,fccn)/quem-somos

- Inlink #1
    - Captured: “2024-03-01 10:00:00”
    - Source: pt,fccn)/
    - Anchor: “Quem somos”

Final output for the next stage: **3 surts**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
    - Captured: “2024-03-01 09:00:00”
    - Outlinks:
        - (pt.fct/,”Home”)
- Inlink #1
    - Captured: “2024-03-01 09:00:00”
    - Source: pt,fct)/
    - Anchor: “Home”
- Inlink #2
    - Captured: “2024-03-01 10:00:00”
    - Source: pt,fccn)/
    - Anchor: “Fundação Ciência Tec.”

pt,fccn)/

- Doc #1 (url: “https://fccn.pt”)
    - Captured: “2024-03-01 10:00:00”
    - Outlinks:
        - (pt.fct/,”Fundação Ciência Tec.”)
        - (pt.fccn/quem-somos,”Quem somos”)

pt,fccn)/quem-somos

- Inlink #1
    - Captured: “2024-03-01 10:00:00”
    - Source: pt,fccn)/
    - Anchor: “Quem somos”

Final output for the next stage: 3 surts, **2 Docs**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
    - Captured: “2024-03-01 09:00:00”
    - Outlinks:
        - (pt.fct/,”Home”)
- Inlink #1
    - Captured: “2024-03-01 09:00:00”
    - Source: pt,fct)/
    - Anchor: “Home”
- Inlink #2
    - Captured: “2024-03-01 10:00:00”
    - Source: pt,fccn)/
    - Anchor: “Fundação Ciência Tec.”

pt,fccn)/

- Doc #1 (url: “https://fccn.pt”)
    - Captured: “2024-03-01 10:00:00”
    - Outlinks:
        - (pt.fct/,”Fundação Ciência Tec.”)
        - (pt.fccn/quem-somos,”Quem somos”)

pt,fccn)/quem-somos

- Inlink #1
    - Captured: “2024-03-01 10:00:00”
    - Source: pt,fccn)/
    - Anchor: “Quem somos”

Final output for the next stage: 3 surts, 2 Docs, **3 Inlinks**

Extracting pages and inlinks

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink #2
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”

pt,fccn)/

- Doc #1 (url: “https://fccn.pt”)
  - Captured: “2024-03-01 10:00:00”
  - Outlinks:
    - (pt.fct/,”Fundação Ciência Tec.”)
    - (pt.fccn/quem-somos,”Quem somos”)

pt,fccn)/quem-somos

- Inlink #1
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Quem somos”

Final output for the next stage: 3 surts, 2 Docs, **3 Inlinks**

FAQ #1

Why SURT?

- It’s a normalized URL that is good enough to do matches. It was also used in the image indexing and… works? No particular opinion here, it just works.

Why don’t do all of this is a single stage?

- When processing records, we only have **write** access to the output HashMap; this means that we can only write Docs and Inlinks. This is what allows Hadoop to process all records in parallel without interdependencies. E.g. on the previous example, the **fccn.pt** and **fct.pt** pages may have been processed in different nodes. Hadoop will ensure that output entries that are accessed by multiple nodes like **pt,fct)/** are grouped and can be processed together at the next stage.

What about duplicates, timestamps…

- That’s on the next stage

Assigning inlinks to pages

The matching period

- I’ve gone with 90 days for now (timespan of 180 days)
  - E.g. **fccn.pt** was captured on **2024-03-01 10:00:00**
    1. Inlinks from **2023-12-02** to **2024-05-30** will be added to that document
- This can be easily changed later

Internal vs. external inlinks

- Internal inlinks must match full domain and subdomain (fully qualified domain name)
  - E.g. for document **fct.pt**
    - i. fct.pt/sobre is an **internal** inlink
    - **ii.** **www**.fct.pt/ is an **internal** inlink (**www** is removed from all SURTs, docs and inlinks!)
    - iii. sobre.fct.pt is an **external** inlink
- Internal and external inlinks are stored separately
  - Both limited to size of 1000 (e.g. maximum would be 1000 internal and 1000 external)
  - Can also be easily tweaked
- Only exact duplicate inlinks are removed
  - Inlink #1
    - i. Captured: “2024-03-01 **09:00:00**”
    - ii. Source: pt,fccn)/
    - iii. Anchor: “Quem somos”
  - Inlink #2
    - i. Captured: “2024-03-01 **10:00:00**”
    - ii. Source: pt,fccn)/
    - iii. Anchor: “Quem somos”
  - Inlink #1 and #2 are different as they have different capture times

Example of assigning inlinks to pages

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink #2
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”

Let’s assign the inlinks from the **pt,fct)/** entry

Example of assigning inlinks to pages

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink #2
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
  - InlinkInternal
  - InlinkExternal

We have one document, let’s “copy” it

Example of assigning inlinks to pages

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink #2
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
  - InlinkInternal
    - Inlink
      - Captured: “2024-03-01 09:00:00”
      - Source: pt,fct)/
      - Anchor: “Home”
  - InlinkExternal

Inlink #1 is internal, add to document InlinkInternal Set

Example of assigning inlinks to pages

pt,fct)/

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
- Inlink #1
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink #2
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”

- Doc #1 (url: “https://www.fct.pt”)
  - Captured: “2024-03-01 09:00:00”
  - Outlinks:
    - (pt.fct/,”Home”)
  - InlinkInternal
    - Inlink
      - Captured: “2024-03-01 09:00:00”
      - Source: pt,fct)/
      - Anchor: “Home”
  - InlinkExternal
    - Inlink
      - Captured: “2024-03-01 10:00:00”
      - Source: pt,fccn)/
      - Anchor: “Fundação Ciência Tec.”

Inlink #2 is external, add to document InlinkExternal Set

Inlink output

Doc #1 (url: “https://www.fct.pt”)

- Captured: “2024-03-01 09:00:00”
- Outlinks:
  - (pt.fct/,”Home”)
- InlinkInternal
  - Inlink
    - Captured: “2024-03-01 09:00:00”
    - Source: pt,fct)/
    - Anchor: “Home”
- InlinkExternal
  - Inlink
    - Captured: “2024-03-01 10:00:00”
    - Source: pt,fccn)/
    - Anchor: “Fundação Ciência Tec.”

https://www.fct.pt:

- Inlink
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”

Inlink output

Doc #1 (url: “https://www.fct.pt”)

- Captured: “2024-03-01 09:00:00”
- Outlinks:
  - (pt.fct/,”Home”)
- InlinkInternal
  - Inlink
    - Captured: “2024-03-01 09:00:00”
    - Source: pt,fct)/
    - Anchor: “Home”
- InlinkExternal
  - Inlink
    - Captured: “2024-03-01 10:00:00”
    - Source: pt,fccn)/
    - Anchor: “Fundação Ciência Tec.”

https://www.fct.pt:

- Inlink
  - Captured: “2024-03-01 09:00:00”
  - Source: pt,fct)/
  - Anchor: “Home”
- Inlink
  - Captured: “2024-03-01 10:00:00”
  - Source: pt,fccn)/
  - Anchor: “Fundação Ciência Tec.”