Introduction

At the recent OpenWayback meeting on 27/04/2016 we opened the discussion of the CDX format. Many interesting views and developments were presented. I would like to offer an alternative use case for the CDX format: a corpus.

The motivation comes from the users, i.e. researchers and scholars who want to do things with the archive. In December 2015 I hosted a 2 days workshop on this particular subject at Aarhus University <URL: http://resaw.eu/events/resaw-technical-meetup/>. It was clear from the work being presented that there is a big need for a common format, the question is: will CDX be it?

What is a corpus and when do we use it?

In my view a corpus is just a fancy word for a collection. But there are differences; whereas a collection is typically created by curators in an archiving institution and considered relatively stable a corpus is created and modified by a researcher working on on a research question. This means:

A corpus can consist of arbitrary objects from the web archive and

- It can be extended
- It can be deleted
- It can be shared
- It can be referenced
- It can be saved for the future (think Data Management)
- It can be used by different tools (for example batch processing tools)

And how does this fit into the discussion of a CDX format?

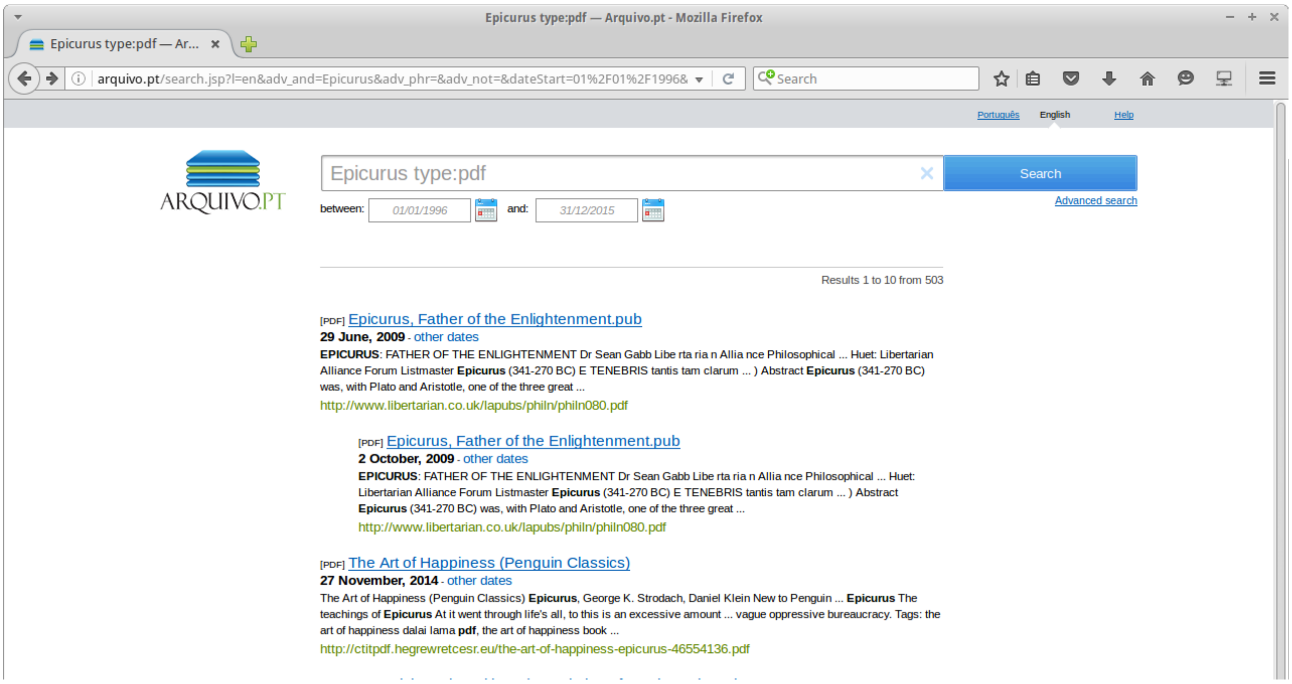
The IIPC is a community effort. Just as we are discussing the specification of WARC we should think thoroughly about the CDX format when we are drafting a specification for archiving institutions to adopt. The same goes for the discussion of an CDX Server API.

CDX is more (or at least, could be more) than an index for OpenWayback. It could be used as a *corpus format*. And the next logical question in this line of thought would be: what can a user do with a corpus?

Here is where it gets interesting. I will give you just a very simple but relevant example that almost all researchers and scholars are doing when working with a web archive. In the following I use the Portuguese Web Archive, Arquivo.pt, as it has a full text search option and is open to the public.

Imagine a philosophy scholar interested in the Greek philosopher “Epicurus”. She sets out to find material in the web archive that she can use in her research.

She enters “Epicurus” in the search field. She is served 68,050 results, most of them web pages (HTML). She decides to filter on file type PDF; the result is now down to 503 documents.


Epicurus type:pdf — Arquivo.pt - Mozilla Firefox
Epicurus type:pdf — Ar...
arquivo.pt/search.jsp?l=en&adv_and=Epicurus&adv_phr=&adv_not=&dateStart=01%2F01%2F1996&
Search
Português
English
Help
ARQUIVO.PT
Search
Advanced search

Epicurus type:pdf

between: 01/01/1996 and: 31/12/2015

Results 1 to 10 from 503

[PDF] Epicurus, Father of the Enlightenment.pub
29 June, 2009 - other dates
EPICURUS: FATHER OF THE ENLIGHTENMENT Dr Sean Gabb Libe rta ria n Allia nce Philosophical ... Huet: Libertarian Alliance Forum Listmaster **Epicurus** (341-270 BC) E TENEBRIS tantis tam clarum ...) Abstract **Epicurus** (341-270 BC) was, with Plato and Aristotle, one of the three great ...
http://www.libertarian.co.uk/lapubs/philn/philn080.pdf

[PDF] Epicurus, Father of the Enlightenment.pub
2 October, 2009 - other dates
EPICURUS: FATHER OF THE ENLIGHTENMENT Dr Sean Gabb Libe rta ria n Allia nce Philosophical ... Huet: Libertarian Alliance Forum Listmaster **Epicurus** (341-270 BC) E TENEBRIS tantis tam clarum ...) Abstract **Epicurus** (341-270 BC) was, with Plato and Aristotle, one of the three great ...
http://www.libertarian.co.uk/lapubs/philn/philn080.pdf

[PDF] The Art of Happiness (Penguin Classics)
27 November, 2014 - other dates
The Art of Happiness (Penguin Classics) **Epicurus**, George K. Strodach, Daniel Klein New to Penguin ... **Epicurus** The teachings of **Epicurus** At it went through life's all, to this is an excessive amount ... vague oppressive bureaucracy. Tags: the art of happiness dalai lama **pdf**, the art of happiness book ...
http://ctitpdf.hegrewretcesr.eu/the-art-of-happiness-epicurus-46554136.pdf

She now goes through the results and selects a total of 42 documents that will make up her corpus at this point. Imagine now that she is able to *export* her results.

What happens from here? The resulting search results basically maps to objects in the archive. Consider an export feature that creates a CDX file to be stored on the server (or even downloaded to the user's PC). This would then make up the research corpus. Because the format is agreed upon, the corpus can be used with tools from the community and the researcher can even donate/deposit her corpus to the archive so that other researchers can use it. This is a way to create customized collections which are basically just “views” on the web archive.

This is just one reason why we need a corpus format.
Below I have listed some other use cases which should be interesting to both researchers and archiving institutions:

- A CDX is a way to “freeze” a search result, something that is inherently difficult in a search engine because the index (and even the engine itself) changes over time.
- A corpus stored in CDX allows for interchange. This would allow researcher B to work on researcher A's corpus for verification or for completely new research. A corpus could be enriched with entries from other archives.
- Because CDX stores a key to each object in the archive it allows for batch processing. This allows a researcher to ask different questions to the same corpus.

What would a new CDX format look like?

A CDX for corpora ***needs to be small***. The good news is that we really only need 3 fields for the lookup key.

A key is composed of:

1. **original URL (non-canonicalized, non-SURTed form)**
2. **timestamp (UTC-format)**
3. **digest (md5/SHA-1)**

Why should we discuss canonicalization?

It is not defined which canonicalization rules we're using. It appears to be only defined in the source code of Heritrix and OpenWayback and even there we currently have no common set of rules in place.

Canonicalization is a tool-specific optimization that changes the crawler's original requested URL. Would we really like to have this in the key?

Are we discussing CDX as an index format?

CDX is not the best index format. Things like SURT and sorting of CDX files are optimizations we do because we all have CDX files. We see institutions doing other kinds of optimization, like the ZipNumCluster format. Why do we try to optimize the format to current tools?
In the lines of the CDX Server API discussion, maybe the question should be: what's the minimum we need? Perhaps we are not asking this because we don't think of CDX as an interchange format.

I have been discussing this thoroughly with my colleague Toke Eskildsen of the Danish Web Archive. Let me end with a line from the resulting blog post:
”It follows easily that the optimal order or even representation of fields depends on tools as well as use case. But how the tools handle CDX data internally really does not matter as long as they expose the CDX Server API correctly and allows for export in an *external* CDX format. The external format should not be dictated by internal use!” <URL: https://sbdevel.wordpress.com/2016/03/18/cdx-musings/>.

Conclusion

Even if your institution is not currently thinking about a corpus format it is likely that some day in the future you will have to consider this. Let's say you have a web archive with replay and search access that you offer to the public. Users will want to:

- *search*
- *filter*
- *select*

and go back to the information they find.

In particular, researchers expect their object of study to behave like the papers, books, photos, papyrus scrolls or whatever objects they usually study -- items which to some degree are stable and can be studied and stored in some way. Currently with web archives, this is difficult. This is where we need a corpus format.