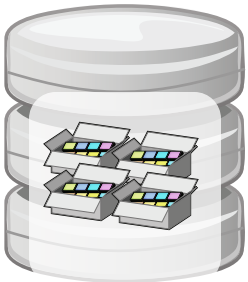
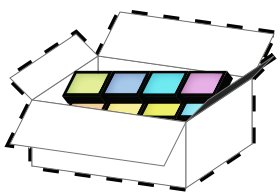
Memory structure: Layers

Layer 4:
User Corpus

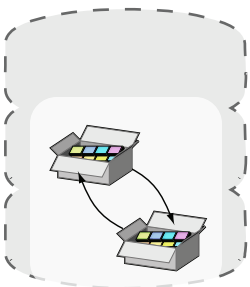


Corpus of WebEntities: A database of WebEntities (**metadata**)

Layer 3:
Dynamic WebEntities Building

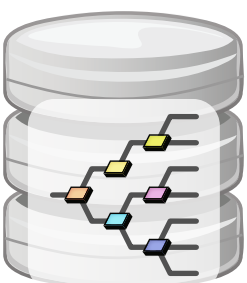


WebEntities: A WebEntity is one or more LRUs. They are **defined in the LRU Index**.

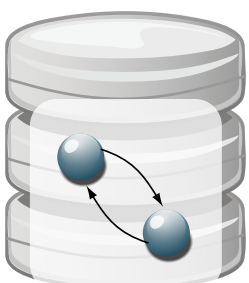


WebEntities Links: **Dynamically build** from LRU Index + Nodes Links

Layer 2:
Nodes & LRU Index



LRU Index: All **LRUs** with **Node** flags and **Pages**



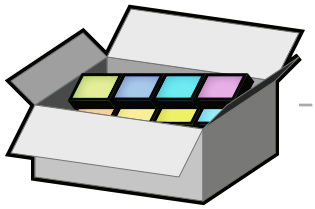
Nodes Links

Layer 1:
RAW Data

Raw Pages: Text files that contain the **links** on each page *(and possibly more)*

SciencesPo. médialab HYPERTEXT CORPUS INITIATIVE

Key Concepts



WebEntity

Typically, it represents a website.
It can be tagged, visualized...
It is the object that the user handles.
It is made of one or several LRUs.

LRU

Also called "Reversed URL",
it corresponds to an URL but is written with the most generic stem first, and the most specific stem last. For example:
".com / google / images".

Stem

It is a component of a LRU.
For us, "images.google.com" is made of the stems images, "google" and ".com".

Web Page

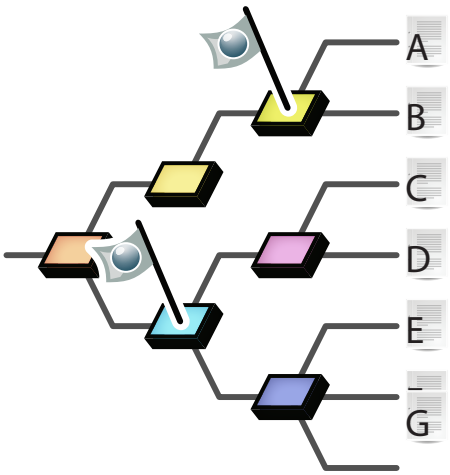
A content accessible in the web through a given URL.
We're mostly interested in links and text.

Link

Web pages have hypertext links (you click on them to browse the web).
We also have links between Nodes, and we have links between WebEntities.

Node

We don't keep all the links between pages in memory (it's too much).
Instead, we keep links between Nodes. A node is just a simplification: close pages are approximated as a single node.



LRU Index

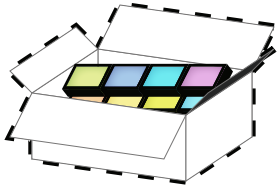
The main index. Can be imagined as a tree of the stems composing the LRUs.
Some LRUs are flagged as Nodes or as WebEntities.

Memory structure: Data Flowing from a Harvest

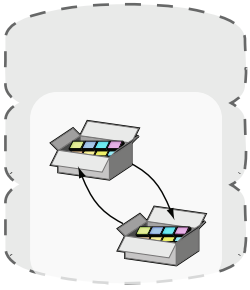
Layer 4:
User Corpus

Corpus of WebEntities

Layer 3:
Dynamic WebEntities Building

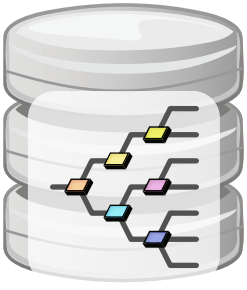


WebEntities



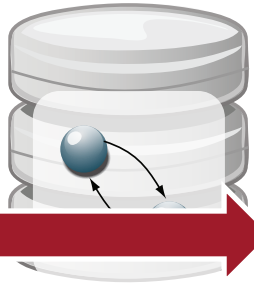
WebEntities Links

Layer 2:
Nodes & LRU Index



LRU Index

3. The LRUs of the Pages, and the corresponding Nodes are stored in the LRU Index.



Nodes Links

4. Links between Pages are agregated as Links between Nodes and stored.

Layer 1:
RAW Data

Raw Pages

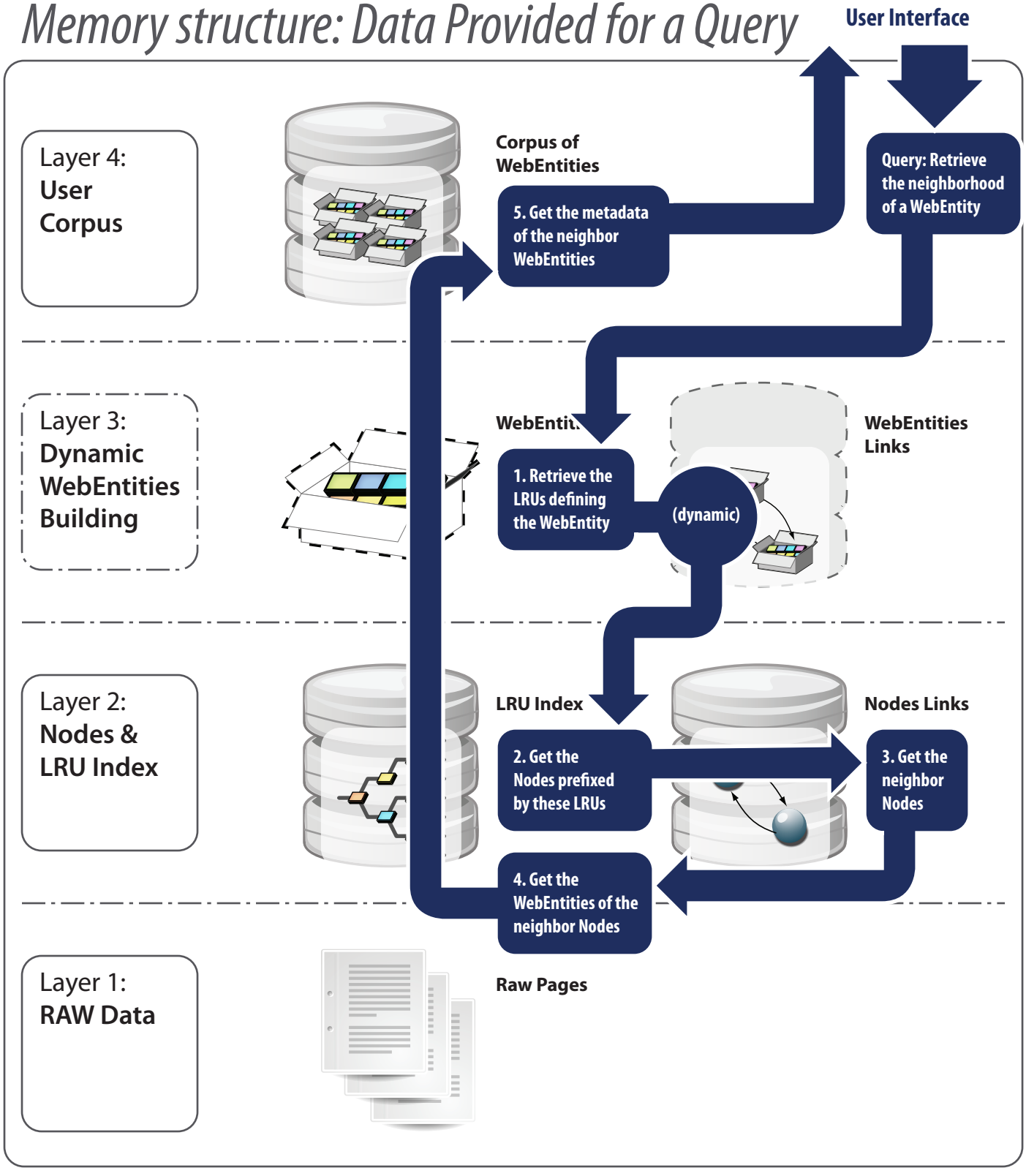
1. Harvest: Raw Data comes in and is stored in the Layer 1.

2. The content added to the RAW Data (Pages) is parsed as LRUs and Nodes.

The Web

SciencesPo. | médialab | HYPERTEXT CORPUS INITIATIVE

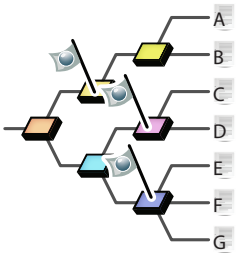
Memory structure: Data Provided for a Query



Process: Retrieving the Links of a WebEntity

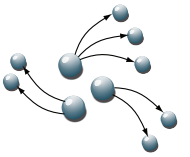
1. Get WebEntity

2. Get the one or several LRUs defining it

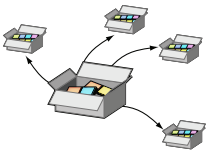


3. From the LRU Index, get all LRUs prefixed by these LRUs (from the WebEntity) that are flagged as Nodes.

4. It defines a certain number of Nodes

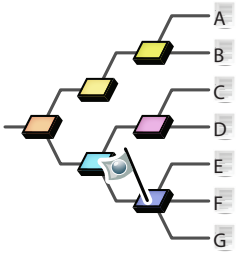


5. Get the Nodes Links from the Layer 2.

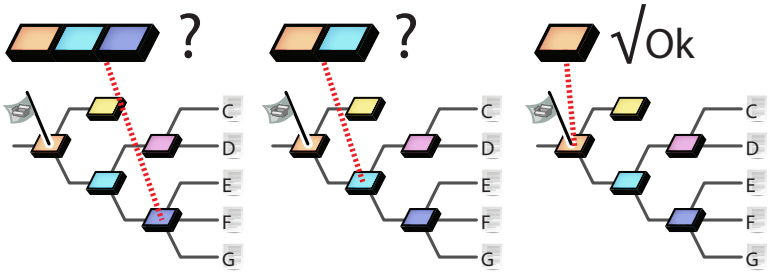


6. From the LRU Index, get the WebEntities of these Nodes (see below).

Process: Getting a WebEntity from a Node



1. In the LRU Index, get the LRU of the Node.

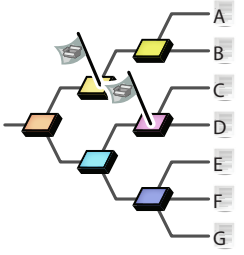


2. Search the smallest sub-LRU wich is a WebEntity
(go back to the root of the tree and get the last WebEntity flag)

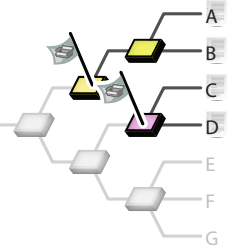
Process: Getting the pages in a WebEntity

1. Get WebEntity

2. Get the one or several LRUs defining it



3. From the LRU Index, get all LRUs prefixed by these LRUs (from the WebEntity) that are flagged as Nodes.



4. In the LRU Index, "walk through the tree" from the LRUs to the last branches that are Pages.

SciencesPo. médialab HYPERTEXT CORPUS INITIATIVE