European Commission Seventh Framework Programme

Call: FP7-ICT-2007-1, Activity: ICT-1-4.1

Contract No: 216267

Coordinated development roadmap

Deliverable No: 10.3

Version 1.0

Editor: Julien Masanès

Work Package: WP10

Status: Version 1.0

Date: 29/1/2009

Dissemination Level: PU

Project Overview

Project Name: LiWA – Living Web Archives

Call Identifier: FP7-ICT-2007-1

Activity Code: ICT-1-4.1

Contract No: 216267

Partners:

1. Coordinator: Universität Hannover, Learning Lab Lower Saxony (L3S), Germany
2. European Archive Foundation (EA), Netherlands
3. Max-Planck-Institut für Informatik (MPG), Germany
4. Computer and Automation Research Institute, Hungarian Academy of Sciences (MTA SZTAKI), Hungary
5. Stichting Nederlands Instituut voor Beeld en Geluid (BeG), Netherlands
6. Hanzo Archives Limited (HANZO), United Kingdom
7. National Library of the Czech Republic (NLP), CZ
8. Moravian Library (MZK), CZ

Document Control

Title: Coordinated development roadmap

Author/Editor: Julien Masanès, Radu Pop

Document History

Version	Date	Author/Editor	Description/Comments
0.1	2008-11-20	Julien Masanès, Radu Pop	First draft. Waiting for input from IIPC
0.2	2009-1-15	Julien Masanès	Second draft including input from IIPC TCO
1.0	2009-1-29	Julien Masanès, Thomas Risse	V1.0 including corrections from Thomas Risse

Legal Notices

Contents

1. EXECUTIVE SUMMARY 5

2. INTRODUCTION 6

3. INTERNATIONAL INTERNET PRESERVATION CONSORTIUM 7

4. WORKING TOGETHER 9

4.1. AREAS OF WORK 10
4.2. PLANNING OF INTERACTION 11

5. LIWA DEVELOPMENT STRATEGY 13

5.1. INTEGRATION CYCLES 14
5.1.1. FIRST INTEGRATION CYCLE 14
5.1.2. SECOND INTEGRATION CYCLE 14
5.2. INTERACTION WITH IIPC DEVELOPMENT 15
5.2.1. CODE LEVEL 15
5.2.2. API LEVEL 15
5.2.3. ARTEFACTS LEVEL 16
5.2.4. EXTERNAL MODULES LEVEL 16

1. Executive summary

In the domain addressed by LiWA, web archiving, the International Internet Preservation Consortium is playing an important role both to organize the community of heritage institutions and to develop open source tools. LiWA has been thought as a virtual 'R&D' extension of this effort since the very beginning.

Although LiWA has its own research agenda, we try our best to position it as useful project for the Web Archiving community and given the prominence of this IIPC in this community, this includes following and coordinating closely where possible with IIPC owns development effort.

The organization of a joint meeting in September 2008 and the preparation of a Joint Workshop in 2009 for developers reflect this willingness of both groups to work together.

This document analyses three important aspects in this regard:

- IIPC achievements, work style and plans.
- Areas of work for both groups and potential interactions
- LiWA integration strategy to adapt its research output to IIPC reference archiving platform.

We provide in this document the current view on this topic, which we will revise next year to adapt to potential change in IIPC plans and revise, if necessary the LiWA strategy in this domain.

2. Introduction

Since 2002-2003, a large effort of joint development of tools has been made in the web archiving community. The formation of IIPC and the start of the open source set of tools including the web archiving crawler Heritrix were important step taken by this community to ensure that:

- All institution would have the tool to develop their own web collections and control the way it is shaped
- Similar approaches, methods and vocabulary would span across the world heritage institutions to accomplish the mission of preserving the web
- Collections would be organized and stored in a compatible manner, ensuring future interoperability and cross access

When this effort started, web technologies were obviously in an early stage with less dynamic sites, videos were not everywhere and web spam was rare. The effort therefore concentrated on archiving basic site with standard type of content. No automatic selection capacity was provided to focus the crawl on important content or at least avoid web spam. Richer media like video was deemed to be secondary.

However, since 2005-2006, limits of this approach have been reached. To illustrate this, let's mention the fact that in a recent web archiving expert session organised by IIPC (IIPC Harvesting Web storming Session, Aarhus, 16th September, 2008) the top three difficulties in archival web crawling mentioned were (in order of citation):

- Advanced web design (JavaScript, Flash, video, Ajax etc.)
- Spam/Traps
- Streaming media

Interestingly, these topics are at the core of LiWA effort. However, these limits are not so easy to address in the current scheme of tools and methods. A research and development effort is necessary to go beyond the current approach, based on sequential, parsing-based crawling. LiWA not only specifically addresses the above-mentioned difficulties but do so in exploring new ways of capturing and organizing web archives.

However, to ensure that the adoption and impact of LiWA research is optimal, it is necessary to decide how best to position LiWA effort in this framework.

3. International Internet Preservation Consortium

The International Internet Preservation Consortium (IIPC) was chartered in July 2003 at the initiative of Internet Archive and Bibliothèque National de France to build a framework of collaboration between Heritage institutions to preserve the memory of the Internet in general and the Web in particular. At that time, initiatives in this domain had been fragmented and many actors felt the need for coordination.

The goals of the consortium were and are still:

- To enable the collection, preservation and long-term access of a rich body of Internet content from around the world.
- To foster the development and use of common tools, techniques and standards for the creation of international archives.
- To be a strong international advocate for initiatives and legislation that encourages the collection, preservation and access to Internet content.
- To encourage and support libraries, archives, museums and cultural heritage institutions everywhere to address Internet content collecting and preservation.

During the first 3 years, the number of participants was limited to the following 12 institutions that had already or were starting web archiving programmes: National libraries of Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, The British Library (UK), The Library of Congress (USA) and the Internet Archive (USA).

In 2007, the consortium was opened to new members and it currently comprises 39 members[1] from all continents.

IIPC is not a legal entity but an agreement to organize collaboration between institutions. The daily operations and the development capacity are provided through good willingness from institutions and a limited budget coming from participation fees (ranging from 2000 to 8000 Euros per year depending on institution's size). This budget is used to support meeting and coordination and to fund development projects. It usually comes in addition to institution funding for project and has helped supporting development of tool like the archiving quality crawler Heritrix[2] developed by Internet Archive or the Warc tools[3] developed by Hanzo Archives and implementing the ISO 28500 WARC format[4].

These projects have been a good way for this community to develop tools they needed, and the provision of a full range of tools from capture to access and search in open source is a great achievement of IIPC. The development style has been collaborative most of time (for instance between Internet Archive and the Nordic Libraries lending some programmers for a period of

[1] See the full list of members at http://www.netpreserve.org/about/members.php (last visited 20/1/2009)

[2] http://crawler.archive.org/ (last visited 20/1/2009)

[3] http://code.google.com/p/warc-tools/ (last visited 20/1/2009)

[4] http://bibnum.bnf.fr/WARC/ (last visited 20/1/2009)

time) but is neither a typical open source project based on individual good willingness and participation, nor a company style mode of development with formal requirement and planning.

In between these two models, IIPC has been able to achieve significant results with loose coordination and no strong steering. The exercise is mainly about leveraging individual institution effort, finding and fostering synergies and factoring resources where possible.

The wide adoptions of tools developed (most heritage institutions engaging with Web Archiving around the world are using IIPC tools) and the adoption by ISO of the IIPC WARC standard prove that this work style has been well fitted to the challenges and the type of partners in this domain, specifically the combination of traditional heritage institutions and new type of foundations like the Internet Archive.

LiWA has to understand this work style to set its expectation at the right level in terms of coordination[5].

[5] The fact that 3 LiWA members are also member of IIPC and that the IIPC technical committee includes Mark Middleton and Julien Masanès (chair) are helpful in this regard.

4. Working together

As seen in the previous section, we are working in a domain where the community is already structured and has already developed a set of tools to enable their practice. The good news is that these tools are open source and therefore easier to understand, improve and complement.

Typically, in such a situation, two options are available: one is to start from scratch, rewrite all the tools and fork from current code base, the other is to develop new tools that complement and augment the capacity of the current ones.

LiWA has chosen the second option for it enables to capitalize on the experience and achievements or the last 5 years. One of the reasons that made this choice possible is the modularity of Heritrix the central tool for web archiving.

However, the main challenge in doing so is to reach reliable and durable articulation of complex pieces of software with new components. And this requires some visibility that both groups need from one another. In this respect, the fact that IIPC is structured as a loose consortium results in a development framework where visibility and planning are not always easy to pin down.

The IIPC consortium chartered specific working groups on different areas of the Web archiving framework. The four working groups are: Standards, Harvesting, Access, and Preservation.

From the LiWA project perspective, the interactions with IIPC development mainly concern the Harvesting working group, focused on the development and improvement of Heritrix crawler, as well as on the support for the WARC file format.

The web group of Internet Archive in charge of Heritrix development is very committed to make the new version of Heritrix (version 2.0 released in 2008) a developer friendly framework so that contributions can be made both within and around the Heritrix code. Actually several of the new features that Heritrix 2.XX provides facilitate the use of Heritrix in the context of various need, providing among other things the ability to sort the frontier (queue of URLs), a more stable interface etc.

A first meeting with both IIPC and LiWA was organized in September 2008 to assess what needs to be done for enabling a durable framework for contributions. The timing is good as, before then, the development of Heritrix focussed mainly on finalizing the continuous crawl version of Heritrix (version 2.n). This version has been released at the end of 2008. IIPC and especially Internet Archive are starting this year to think about Heritrix Future evolutions. For this purpose, it was decided that a full workshop will be organized in 2009 with Heritrix development team and advanced users and developers, among which LiWA partners, to assess new directions and priority, as well as to better coordinate exterior contributions like LiWA's.

In the meantime, LiWA team has had time to get familiar with Heritrix structure and code and will be in situation of defining what evolution/change they need, if they need some.

Of course, both LiWA and IIPC remain separate groups with separate agendas. There will be no alignment of objectives and means and resources will not be factored. However, building on existing blocks and working together to make this foundations open for separate but compatible work is the basis of this approach.

In parallel, some LiWA partners, will also build on top of their own platform to leverage LiWA research. This is the case for instance of Hanzo Archives who uses their own platform to apply and test LiWA research output.

4.1. Areas of work

The coordination effort aims at complementing each partner's work where possible. For this, objectives that both IIPC and LiWA are pursuing and potential duplicate work or overlap that they might have need to be assess.

To help in this, we have drafted a map of current development plans for both organizations. This roadmap is not constraining in any ways for both LiWA and IIPC but rather should serve as a tool for LiWA to better coordinate with the main user community in this field.

The following diagram (Fig. 1) presents areas of work of both groups (as documented end of 2008) and highlights the potential interactions.

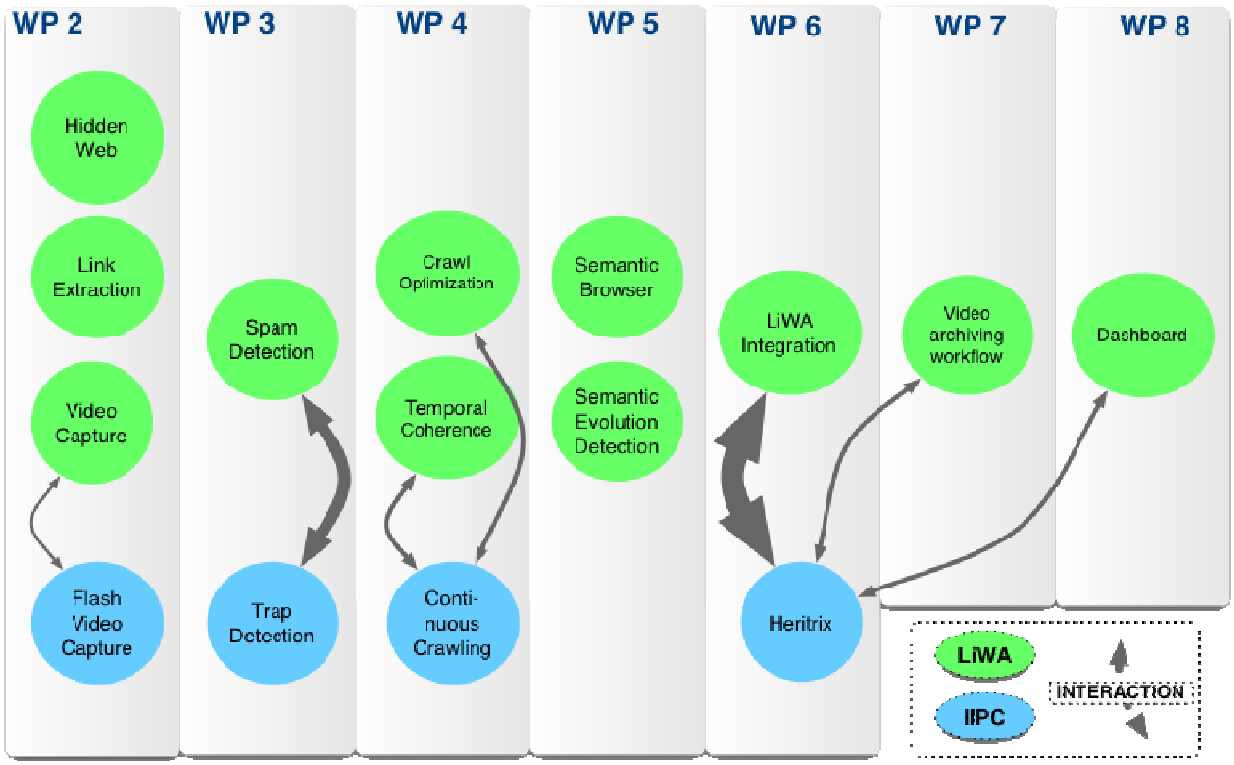

Figure 1: LiWA IIPC current area of work

For **WP 2** (archive completeness) there is no research/experimentation going on in IIPC, beyond the continuous improvement of Heritrix, except one module that is been developed to archive flash embedded video, which complements well the work LiWA is doing on streaming archiving. LiWA is focussing on non-http protocol, whereas Flash uses the http protocol.

For **WP3** (and **WP10** as it is related to the future web archiving platform), IIPC is going to start working on a trap detection knowledge base, to enable crawl engineers to share patterns and settings to detect and avoid crawler traps. This approach (leveraging the community input to share it) is similar to the one that we envisage for Spam detection, and some of it could overlap. Although the spam detection is based on a rich set of features, some of the method used could be shared and we therefore envisioned that a potential rich interaction occurs here.

WP4 is working on temporal aspect of crawl and the current work done by Internet Archive on behalf of the IIPC on continuous crawling is closely related as it is based on assessment of change frequencies as well. However, the focus of **WP4** is currently on smaller crawl and focusing on quality and experience of the user (avoiding time drifting in web archives). Future work however, could be used for larger crawls (like national domain crawls) and linking with Heritrix development work on continuous crawl is for this desirable.

WP5 is exploring a domain which is discussed but where no work is started or planned at this moment within IIPC.

WP6 is the key WP where interaction will take place, as modules are being integrated and tested on top of Heritrix developed by IIPC. This integration work is carried out by the European Archive foundation that is also members of IIPC (and represented in the IIPC technical committee).

WP7 and **WP8** (applications work packages of LiWA) will also be connected to the Heritrix work stream as they use it as a central tool. However, this is through the integration work carried on by **WP6**.

As this analysis shows, there is little if any overlap with IIPC current work agenda. IIPC is not launching any project in the areas that have been identified by LiWA as research topics. Main recent projects launched by IIPC are related to Heritrix (already mentioned) and implementation of the WARC ISO standard (WARC tools).

There is therefore the possibility to maximise complementarily between the two efforts.

4.2. Planning of interaction

Organizing fruitful interactions between the two communities (researchers and practitioners) benefit from physical meetings. However, budget constraints limit what can be done in this regard and we have tried to identify the best moment and, where possible combine with other events.

The key milestones for LiWA are the technology delivery and integration. The integration will be done in two iterations.

For ensuring we have both a clear view on the LiWA development process and practice, but not too late to interact with IIPC, we have scheduled the first meeting during IWAW/ECDL in September 2008 in Aarhus, in the middle of LiWA's first year.

We plan to participate to the Heritrix advanced users workshop mid 2009, where potential new orientation for Heritrix development and assessment of developer needs will be made.

Finally, a third meeting will be scheduled, before the final integration prototype is finished.

It is worth noting that the core user group, whom members are almost all part of IIPC, will have it's user-centered activity running in parallel to this coordination of development efforts. We expect this group to feedback on user needs and tests where the meeting described below are centered on the development coordination.

5. LiWA Development Strategy

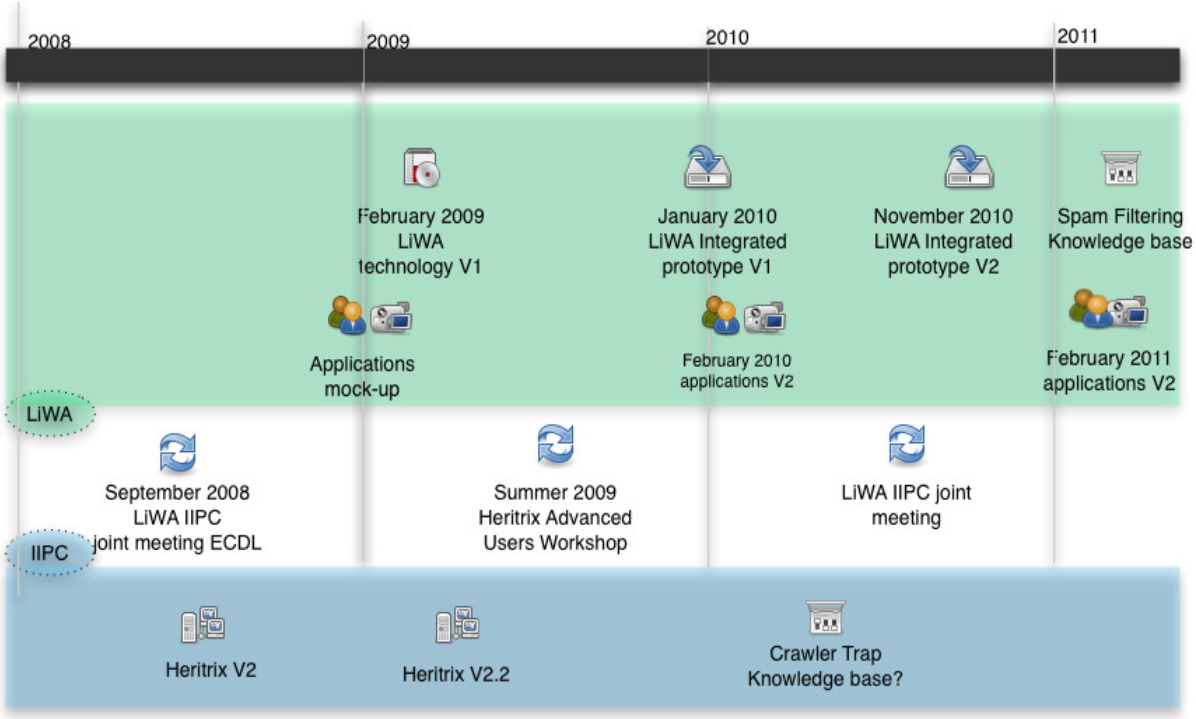

Figure 2: LiWA IIPC planning of interaction

Our development strategy involves 2 integration cycles, as illustrated in the figure above (Fig. 2). The major milestones of the project correspond to the 2 releases of the LiWA prototype:

- M6.1 (T23) – marks the end of the first integration cycle. The first version of the prototype includes the first version of LiWA new technologies.
- M6.2 (T33) – marks the end of the second integration cycle and corresponds to the second release of LiWA prototype.

The test activities during the two cycles will be correlated with the IIPC roadmap and with the latest releases of IIPC tools (e.g. Heritrix roadmap, WARC Tools, etc.)

At time T12 (January 2009), the first release of LiWA's new technologies will be available and we will start the integration tests:

- Enhanced Capture Technology - WP2
- Archive Filtering Technology - WP3
- Coherence Analysis Technology - WP4
- Terminology Extraction Technology - WP5

5.1. Integration cycles

5.1.1. First integration cycle

In the first integration cycle of 11 months (from T12 to T23) we will perform the first round of integration tests, organised in the following main topics:

Scripts to automatically launch the two crawlers: Heritrix and Hanzo Crawler

We will create the appropriate framework to run Heritrix crawler in two different versions: the latest version of Heritrix (2.0.2), as provided by IIPC, and a LiWA modified version of Heritrix. The modified version mainly includes the LiWA extensions to analyse the time coherence: an extension of the "Frontier" object, developed by Max-Plank-Institut to gather supplementary information at crawl time. All this data is stored in an auxiliary database in order to facilitate the study of the time coherence. Other LiWA plug-ins (e.g. the Spam filter module) are planned to be included in the modified version.

We use a predefined set of data for the test crawls, based on our prior experience with large and complex Web sites (e.g. www.mod.uk, www.cabinetoffice.gov.uk, www.royalnavy.mod.uk). The scripts will launch both versions of the crawler with the same list of seeds. The goal is to provide the same test environment and to compare the crawl results obtained with the two versions of Heritrix.

Deployment of the external modules

LiWA external modules need specific libraries and other software sub-modules to run properly. For instance, the Spam filter module uses an external spam engine and its libraries to identify and analyse features in the Web pages. The time coherence analyser uses a local database to store the crawl-time information.

All these software components must be deployed in the test-bed framework and installed together with the LiWA modified version of Heritrix.

The connectivity between different LiWA modules is based on Web services communication. Our test scenarios involve the deployment and run of an Apache Web server on a dedicated LiWA server. The test-bed framework must provide a seamless communication and scaling test must be also performed.

At time T23 (December 2009), the second release of LiWA's new technologies will be available for testing. This represents also the intermediary milestone for LiWA project. Based on the experience gained through the first round of tests, we will refine the data sets and the description of the test gardens. Moreover, the synthetic data sets will be available for testing.

This is also an appropriate moment to coordinate with IIPC roadmap and to analyse the latest releases of the tools.

5.1.2. Second integration cycle

In the second integration cycle of 10 months (from T23 to T33) we will continue the integration tests on the second version of LiWA's new technologies. Using the same framework defined for

the first cycle, we will perform distinct sets of tests for the two infrastructures: European Archive and Hanzo Archives.

5.2. Interaction with IIPC development

As we described in the previous section, the first important synchronization point between LiWA results and IIPC tools is marked by the first milestone (M6.1) of the project, scheduled for the end of December 2009 (T23). At the current stage of the project, we stated the problems to tackle, we defined the complete list of requirements, and each research topic advanced on its state-of-the-art evaluation and on possible solutions to explore. During the first integration cycle in LiWA we will test and evaluate the new technologies developed by the four research areas of the project.

Our goal is to show at the end of this integration cycle that LiWA's new technologies effectively improve the current practices in Web harvesting and to propose to the IIPC consortium the adoption of our results. At the same time, the interactions with IIPC group will go in both ways, in the sense that LiWA development will be kept up to date with the latest releases of Heritrix.

Looking more precisely into the tools development, we identified three possible levels of interaction between LiWA project and the IIPC “Harvesting” working group:

- Source code level
- Generic API level
- New internal modules (artefacts) level
- New external modules level

5.2.1. Code level

This level of interaction implies the direct access and updates on the source code of Heritrix. From LiWA perspective, the research areas working at this level are mainly the capturing enhancements and the temporal coherence analysis. In a later phase, we might consider also the spam filtering techniques acting at this level.

For example, for the temporal coherence analysis (WP4) we extended the “Frontier” class in Heritrix, by inserting some supplementary methods to capture and store additional information during the crawl. At this stage of our research we perform a temporal analysis based on the current crawls and we focus on developing efficient algorithms to (partially) solve the temporal incoherence.

Our goal for the first version of the coherence analysis technology is to develop an enhanced version of the code, which will include the implementation of our algorithms. After the first test cycle and the assessment of the results, we will propose to the IIPC technical committee the integration of this version in the source code of Heritrix.

5.2.2. API level

This level of interaction involves the definition of an extended API that can be used by the LiWA new modules to interact with Heritrix. This extended API includes the different APIs declared by

each LiWA module, but it also includes some new methods declared for Heritrix and required by LiWA modules.

For instance, the API defined for the spam filter module includes several specific methods, like: SpamFeaturesFromPage(ProcessorURI), GraphFromPageFeatures(List <PageFeature>), HostFeatureFromPage(List <PageFeature>), SpamModel(List <HostFeature>), analyse(crawlID, List <newWARCFiles>), etc. On the other hand, the spam filter module will need some specific input from Heritrix for the features extraction. The corresponding methods required by this module are still to be defined and to be proposed as an API extension for Heritrix.

5.2.3. Artefacts level

This represents the “lightest” interaction level between Heritrix and the LiWA modules. Different internal modules for Heritrix developed in LiWA (e.g. the crawl-time coherence analyser, the crawl-time spam filter) will work like plug-ins that can be activated and deactivated from the parameters configuration. They will not interact with the source code development of Heritrix and they will not impact on the performance of the crawler when deactivated.

These internal modules can be selected and added in the extractor’s pipeline of the crawler.

5.2.4. External modules level

This interaction level mainly concerns the post-processing stage of Web archiving. The research areas dedicated to this stage are the off-line coherence analysis, off-line spam filtering and the terminology extraction.

The external modules developed in LiWA will interact with different other tools in IIPC chain, like for instance the WARC tools.

After the first test cycle, we will evaluate and assess the improvements of the archive's quality brought by LiWA tools and we will propose the adoption of these external processors in the IIPC's toolkit.