
**Dipartimento di Ingegneria e Scienza dell'Informazione**

- KnowDive Group -

# Crawling Metadata

| Document data: | Reference persons: |
| --- | --- |
| 09.03.24.v01.r02 | Ilya Zaihrayeu {ilya@disi.unitn.it}<br>Denys Babenko {babenko@disi.unitn.it} |

**Abstract:** The user's information space, together the desktop and the web, is a huge container of interlinked metadata-rich artifacts. Nowadays systems managing such artifacts are either domain specific or do not use the potential of the metadata. To address the problem various metadata systems are developed. They all have to deal with common crucial aspect which is metadata extraction and integration. Current thesis aims at building a bridge between the user's information space and a large scale metadata management system, by developing a framework for artifact crawling as well as extraction and integration of metadata.

**Keywords:** metadata, crawling, adapter, integration

**Objectives:** the current thesis aims at:

- definition of metadata crawlers (for example, crawlers for Facebook, Flickr, Gmail, Outlook, Desktop, etc.)
- definition of metadata integrators (for example, Facebook and Flickr integrator)
- implementation of crawlers, adapters, integrators
- integration of developed software within a large scale metadata system (Sweb[1])

**Work plan:** the effort required to deliver the current thesis is estimated as a 7 months full time activity of a master student. The thesis work plan is presented below. The delivery dates are rather indicative than compulsory, however, the student is highly encouraged to follow the proposed timeline and s/he should notify the thesis advisor whenever a deadline cannot be met. In the list below, **M** stands for "month", e.g., **M1** is the end of the first month after the thesis starting date.

**M0.5:** report on types of artifacts and their metadata in state of the art systems
**M1.5:** report on state of the art and metadata management systems approaches to metadata crawling, extraction and integration
**M2.5:** report on the definition of the framework for metadata extraction;
**M2.75:** report on implementation and integration requirement analysis;
**M5.75:** implementation and integration of developed framework;
**M6:** report on testing results of implemented framework;
**M6.5**: revised and final implementation based on the results of the evaluation;
**M7**: final modifications and a thesis report, which will summarize the completed work.

**Requirements:** the candidate student for this thesis should meet the following requirements:

- bachelor degree in computer science or in a related field;
- good knowledge of OOP and of the Java programming language. Programming experience in large projects is a plus;
- good knowledge of relational databases and related technologies. Particularly, the candidate should be familiar with the principles of database design (e.g., normal forms), with SQL primitives, with the entity-relationship model, with how to access and manipulate database objects from Java using a JDBC driver.

[1] Sweb system, see http://dit.unitn.it/~knowdive.