Disseminating Heritage Records as Linked Open Data

Sam Coppens, Erik Mannens and Rik Van de Walle

Department of Electronics and Information Systems – Multimedia Lab Ghent University – IBBT

***Abstract*—In Flanders, Belgium, many heritage institutions disseminate their metadata using the Open Archives Initiative Protocol for Metadata Harvesting. However this protocol does not offer granular access to the metadata. This can be solved by exposing the metadata as Linked Data. For this, we developed a semantic metadata schema consisting of two layers: One layer gives a Dublin Core description and is responsible for searching the whole dataset. The other layer holds a reference to the original metadata record, e.g., MARC record. Doing this, the user can still access the original record, once he found the data of interest using the Dublin Core description. Then, these metadata records are enriched in two stages: First, we enrich the records internally, interlinking all the harvested metadata from the Flemish heritage institutions. Then, we enrich the records with other datasets like DBpedia, weaving the information into the Web of Data. For publishing the records as linked open data, we enhanced the OAI2LOD Server to import data coming from different OAI-PMH repositories, to expose the records as linked open data using our developed metadata schema and to enrich the records using our metadata enrichment algorithm. This way, the data from Flemish heritage repositories are linked with each other and published as linked open data.**

***Index Terms*—Heritage, Linked Open Data, Metadata Schema, Multimedia Database, Semantic Web Technologies**

I. INTRODUCTION

A frequently heard criticism on the use of Information and communication technologies (ICT) within a heritage context is the individualization of the experience. In Flanders, Belgium, the project Heritage2.0 [1] aims at building a framework for augmenting the social interactive location-based heritage perception within a network of heritage sites by means of mobile devices, i.e., Personal Digital Assistants (PDAs).

These mobile devices guide the participants through the heritage site and provide the visitor with the appropriate information, based on their location within the heritage site. The provided information comes not only from the visited heritage institution, but also from other Flemish heritage institutions.

Many institutions grant access to their repositories via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, [2]). This protocol has some drawbacks, e.g., the resources are not accessible via dereferencable URIs. To overcome these issues, the metadata coming from the OAI-PMH endpoints, is published according to the principles of Linked Data (see Section 5, [3]) and access to the metadata is provided by a SPARQL Protocol and RDF Query Language (SPARQL, [4]) endpoint.

To handle the plurality of metadata schemes used by the heritage institutions, all the information of the heritage institutions have to be disseminated on an interoperable manner. For this, we developed a semantic metadata schema for the exchange and publishing of the heritage information. This information is then weaved into the Web of data via metadata enrichments with information coming from other datasets, e.g., DBpedia [5], GeoNames [6], or other heritage institutions.

II. OAI-PMH

OAI-PMH is a protocol used for sharing and exchanging metadata. This protocol is very popular in the domain of digital libraries. Currently more than 1700 repositories expose their metadata descriptions for several millions of items via the OAI-PMH protocol [7].

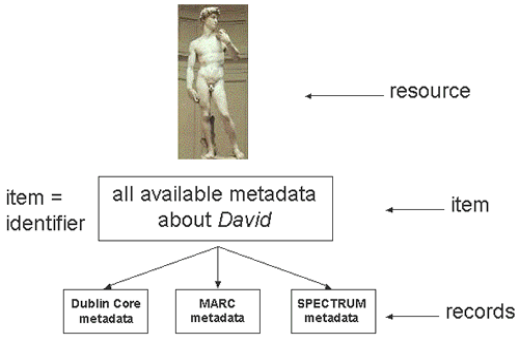

Fig. 1. Basic concepts of OAI-PMH

Client applications can use this protocol to harvest metadata coming from Data Providers using open standards like Hypertext Transfer Protocol (HTTP, [8]), and Extensible Markup Language (XML, [9]). By implementing wrappers on top of their metadata repositories, institutions can easily expose their metadata via OAI-PMH.

The number of OAI-PMH endpoints is expected to grow, because many popular open source digital library systems, such as Fedora [10], DSpace [11], and EPrints [12], provide an OAI-PMH endpoint by default. Another reason for the growth of these OAI-PMH endpoints, is that major attempts to build union catalogues, e.g., The European Library project [13], rely on this protocol for indexing metadata originating

Manuscript received on March 5, 2008.
E-mail: sam.coppens@ugent.be.

from different remote sources.

2.1 Principles of OAI-PMH

The OAI-PMH protocol disseminates the metadata about the items of a repository. These items describe digital or non-digital resources. An item is identified by a URI. Each item can have multiple metadata records. Each record is described by a certain metadata schema. Thus, each item can have multiple records, each described by a different metadata schema. These schemes are chosen by the data provider to suit their domain-specific demands. The most frequently used schemes are RFC1807 [14], MARC [15], MARC-21 [16], MODS [17], and METS [18]. To guarantee a basic level of interoperability one of those metadata schemes must be unqualified Dublin Core. Figure 1 shows the basic concepts of the protocol.

The OAI-PMH protocol is based on HTTP. It supports six request types ("verbs"). The request arguments are issued as GET or POST parameters. The responses are encoded in XML syntax. This functioning is shown in Figure 2.

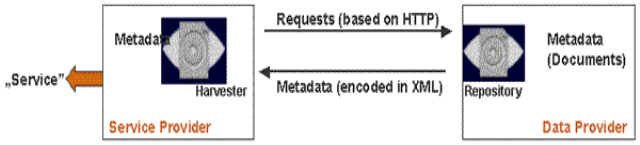

Fig. 2. Basic functioning of OAI-PMH

As already mentioned, there are six verbs, or request types, defined in the OAI-PMH. The Identify request retrieves administrative metadata about the repository, e.g., the name or owner of the repository. GetRecord retrieves metadata about an item in a certain metadataformat. ListRecords harvests all metadata records in a certain metadata format for all items in the repository. ListIdentifiers lists all the identifiers of the available items. ListMetadataFormats returns the available metadata formats used in the repository. Finally, ListSets gives the available sets in the OAI-PMH repository. Figure 3 gives an example of an OAI-PMH request with the XML-encoded response.

2.2 Limitations

The OAI-PMH protocol is a widespread protocol for disseminating metadata records. It is aimed at sharing metadata records over HTTP. Though, the protocol has some limitations.

A first limitation is the use of non-dereferencable identities. To retrieve information from a OAI-PMH repository, a client must execute a HTTP GET request on an OAI-PMH specific URI (see Figure 2). Web clients that do not know this protocol, cannot access this information.

Another limitation of the OAI-PMH protocol are the selection criteria. The client has only limited access to the metadata. The criteria to retrieve a record are the item identifier, the metadata formats, the sets, and the record creation date intervals. Retrieving a record that fulfills a certain condition, except the ones mentioned above, is not possible. For example, a request asking all the records about "the paintings of Pieter Paul Rubens" in the repository is not possible.

Publishing these records according to the principles of Linked Open Data (see Section 5, [3]) and providing a SPARQL endpoint overcomes these problems. In order to search the whole repository, a common metadata schema has to be provided.

a) Request

http://arXiv.org/oai2?
verb=GetRecord
&identifier=oai:arXiv.org:cs/0112017
&metadataPrefix=oai_dc

b) Response

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH
   Xmlns=…>
  <responseDate>2002-02-08T08:55:46Z
  </responseDate>
  <request verb="GetRecord"
   identifier="oai:sample.org:sampleset/
   0112017"
   metadataPrefix="oai_dc">
   http://sample.org/oai
  </request>
  <GetRecord>
   <record>
    <header>
      <identifier>
       oai:sample.org:sampleset/0112017
      </identifier>
      <datestamp>
      2009-04-01
      </datestamp>
      <setSpec>
      Sampleset
      </setSpec>
    </header>
    <metadata>
      <oai_dc:dc
       xmlns:oai_dc= …>
        <dc:title>
         Example OAI-PMH Response
        </dc:title>
        <dc:creator>
         S.Coppens
        </dc:creator>
        <dc:subject>
         Digital Libraries
        </dc:subject>
        <dc:description>
         …
        </dc:description>
        <dc:date>
         2009-04-01
        </dc:date>
      </oai_dc:dc>
    </metadata>
    <about>
    …
    </about>
  </record>
 </GetRecord>
</OAI-PMH>
```

Fig. 3. OAI-PMH request and response example

III. METADATA SCHEMA

When storing digital cultural heritage coming from different institutions the repository has to handle a plurality of descriptive metadata formats. Each specific application area determines which descriptive metadata are necessary. Digital images coming from a library may represent a scanned book. Images coming from a museum may represent an artwork. Both images need other descriptive metadata fields. The digital archive has to be able to search the whole dataset, with data from different data providers. It needs a covering metadata schema for this purpose.

The choice of this covering metadata schema is a non trivial task. Many of the institutions already have descriptive metadata. Are these descriptive metadata stored as metadata or as data? Both strategies have their advantages and disadvantages. When archiving these descriptions as metadata, the repository has to provide a covering metadata schema. The metadata schemes used for the descriptions are very domain-specific. To store the descriptive metadata lossless, the descriptive metadata schema used by the repository should be some kind of smallest common multiple of all the descriptive metadata schemes offered by the institutions. This would be a huge metadata schema, impossible to maintain. That is why a reference to the original metadata, e.g., MARC, is stored, so there is no information loss. On top of this reference, the archive offers a broadly accepted descriptive metadata schema. This schema should be the greatest common divisor of all the metadata schemes offered by the institutions. This gives the repository the necessary tools to search the whole dataset. When finding the data of interest, the original metadata that is referenced can still be presented to the users.

Therefore, we developed a metadata schema consisting of two layers. One layer offers the descriptive metadata, while the other layer stores the reference to the original metadata, which is stored at the institution. For the top layer an OWL DL [19] representation of Dublin Core is chosen. For the reference to the original metadata, we developed an OWL DL provenance schema, indicating from which record the Dublin Core record originates from. This way our metadata schema offers the tools to search the whole dataset of the repository and a reference to the original metadata record for those who want a more detailed description.

3.1 Layered OWL DL Schema

For the definition of new metadata schema we will use the OWL ontology language. The expressiveness of OWL allows us to create fine-grained property definitions by splitting the definition of properties into datatype properties and object properties. A datatype property can take typed literals as value whereas an object property can link to other resources like content items taken from an ontology.

The sublanguage is OWL DL, not OWL FULL. OWL FULL gives the most expressiveness, but does not guarantee the support of reasoning software, while OWL DL is a little less expressive, but it is guaranteed to be completely supported by the reasoners.

3.2 Dublin Core OWL

Descriptive metadata describe the content of the data: subject, author, date of creation, file format, etc. This metadata makes it possible to manage and search the complete repository.

Dublin Core was chosen to describe this top layer of descriptive metadata. Dublin Core is a broadly accepted descriptive schema. The power of this schema is its simplicity and generality. It consists of fifteen fields, a.o., creator, subject, coverage, description, and date. It can answer to the basic questions: Who, What, Where, and When. All the fields in Dublin Core are optional and repeatable. This makes it possible to map almost all the descriptive metadata schemes easily to Dublin Core. This part is automatically offered by the repositories of the institutions disseminating their metadata with the OAI-PMH protocol.

3.3 Provenance OWL

The provenance layer is responsible for delivering a handle to the original record, where the Dublin Core description originates from. For this, the provenance part should deliver at least three things: the metadata namespace of the originating record, the URI of the repository it comes from and the identifier of that originating record in that repository. This part is based on a schema that is used by the OAI-PMH protocol for indicating the provenance of a record. In most of the cases, the Dublin Core description, disseminated by an OAI-PMH repository, is derived via a mapping from another, richer record, e.g., a MARC record. The provenance of such record can be included in the OAI-PMH response within the about container, as depicted in Figure 3. This schema is described via XML schema, which we implemented as an OWL DL ontology.

The XML schema defines a provenance container consisting of a sequence of *originDescription* elements that identify the provenance of the metadata record. Each *originDescription* contains the following information:

- *baseURL*: The base URL of the originating repository from which the metadata record was harvested.
- *identifier*: The unique identifier of the item in the originating repository from which the metadata record was disseminated.
- *datestamp*: The datestamp of that metadata record.
- *metadataNamespace*: The namespace URI of the metadata format of the original record.
- *originDescription*: An optional *originDescription* block which was that obtained when the metadata record was harvested. A set of nested *originDescription* blocks describe provenance over a sequence of harvests.

Each originDescription must also have the following two attributes:

- *harvestDate*: The responsdate of the OAI-PMH response that resulted in the record being harvested from the originating repository.
- *altered*: a Boolean value indicating if the harvested record was altered before being disseminated again.

For the OWL DL description of this schema, a class is made up, *provenanceType*. An object property is defined on this class. The range of this object property is the class *originDescriptionType*. This object property has a minimum cardinality of one. This means that an instance of *provenanceType* holds at least one instance of *originDescriptionType*. This models the sequence of *originDescription* elements as described by the XML schema of the provenance.

The class *originDescriptionType* has six datatype properties: *baseURL*, *identifier* and *metadataNamespace*, which all have a URI as range, *datestamp* and *harvestDate*, which have a string as range, and finally *altered*, which has a Boolean as range. All these six datatype properties are required and have a cardinality of one.

The class *originDescriptionType* has one object property, *originDescription*, which relates an instance of *originDescriptionType* to another instance of *originDescriptionType*. This property is optional, so it has a maximum cardinality of one.

IV. LINKED OPEN DATA SERVER

Now that we have our semantic metadata schema, we can start publishing the OAI-PMH disseminated records of the heritage institutions according to the principles of linked open data (see Section 5, [3]), using our developed metadata schema.

For this we rely on a tool, called OAI2LOD Server [20]. It is a stand-alone server implemented in Java and based on the architecture of the D2RQ Server [21]. The server harvests the metadata described in Dublin Core from a given OAI-PMH endpoint. These records are transformed into RDF/XML [22] and stored in-memory. The metadata are then exposed to various kinds of clients. The server analyses the accept property in the HTTP headers and delivers the metadata either in RDF/XML or in XHTML using the HTTP 303 See Other response. One of the limitations of OAI-PMH was the fact that it uses non-resolvable URNs to identify items. The OAI2LOD Server solves this by offering HTTP URLs.

The second limitation of OAI-PMH was the restricted access to the metadata, as mentioned earlier. The OAI2LOD Server offers a SPARQL endpoint, which gives full access to the stored metadata. With this endpoint it is possible, e.g., to give all the records about "the paintings of Pieter Paul Rubens" in the repository.

The OAI2LOD Server has some drawbacks, though. It serves records from an in-memory Jena [23] RDF model. The number of records a server can host, depends on the amount of memory assigned to the Java Virtual Machine. To overcome this problem, we adapted the OAI2LOD Server to serve records from an Openlink's Virtuoso triple store [24].

Another limitation of OAI2LOD Server is that it can only serve records from one OAI-PMH repository. In order to solve this, we extended the OAI2LOD Server to expose not only sets and records from an OAI-PMH repository, but also collections, which collect the sets and records from a certain OAI-PMH repository. This way, the OAI2LOD Server can harvest records from different OAI-PMH endpoints.

Finally, we enhanced the OAI2LOD Server to serve records not only in Dublin Core, but also in our developed metadata schema consisting of Dublin Core OWL and Provenance OWL. Now, the harvested, transformed metadata records can be exposed as linked open data.

V. METADATA ENRICHMENTS

When publishing metadata the Linked Open Data way, there are four conditions, which have to be fulfilled: The first rule says that all the things offered by the repository should have URIs. The second rule implies that all the URIs that identify things, should have resolvable HTTP URIs. The third rule proposes to deliver useful information whenever a URI is dereferenced. Finally, the last rule recommends that the metadata records should contain links to other related datasets. The first three rules are already covered. The last rule is only partially covered. For now the metadata records hold only a reference to the original record, where the Dublin Core description originates from. This reference is stored in the Provenance class of the metadata record. To weave the metadata properly into the Web of data, we need more metadata enrichments.

The OAI2LOD Server offers the possibility to enrich the metadata with external datasets, but this enrichment does not work well and is extremely inefficient.

Our enhanced enrichment algorithm consists of two stages. During the first stage the records are enriched with records stored in the repository. This way, the records coming from different heritage institutions are enriched with each other. This links all the data coming from the Flemish heritage institutions.

During the second stage, we start enriching the records with datasets from DBpedia and GeoNames. This way the metadata records are weaved into the Web of data.

5.1 Metadata enrichments first stage

For this internal interlinking, similarities between the already harvested and exposed metadata records are looked for. If there is a similarity, the records are enriched with a HTTP link to the related record.

For the similarities, the records are iterated for identifying similarities between values (objects) of certain properties (predicates) of the stored metadata records. We look at similarities for the following Dublin Core properties: title, subject, creator, publisher, and contributor. When a similarity between two records is detected, the records are enriched with each other's HTTP URI through the RDFS (Resource Description Framework Schema, [25]) seeAlso property.

This way, the records coming from the different heritage institutions become interlinked and semantically richer.

5.2 Metadata enrichments second stage

For the external interlinking, two datasets are queried: DBpedia and GeoNames. If they have records with relevant information, the HTTP URIs of these records are added to our stored metadata records through the same RDFS seeAlso property. This is done by querying these two datasets for the values of the following Dublin Core properties of the stored

metadata records: title, subject, creator, publisher, contributor, and coverage.

Now, the metadata records of our linked open data server are linked with the datasets of DBpedia and GeoNames, weaving the records into the Web of data.

VI. CONCLUSION

In this paper, we proposed a way of publishing metadata originating from different heritage institutions according to the linked open data principles.

First, we start from OAI-PMH endpoints, which many institutions already have in place. This protocol disseminates the records at least in the Dublin Core format. This protocol has some limitations, which are the lack of dereferencable HTTP URIs and the restricted access to the metadata. These limitations can be solved by publishing the data as linked open data.

Before publishing the records, a common metadata format has to be defined. For this we developed a layered semantic metadata schema. For the description of the schema, the ontology language OWL DL was chosen. The first layer exposes the metadata in the unqualified Dublin Core format. This way, the whole dataset can be managed and searched. The second layer stores a reference to the original record, where the Dublin Core description comes from. By offering a reference to the original data, the original metadata can still be presented to the user, when findng the data of interest via the first Dublin Core layer.

For publishing the metadata in the developed metadata schema, the OAI2LOD Server is used and enhanced. The shortcomings of this server that have been resolved, are: publishing data coming from different OAI-PMH repositories, adapting the server to serve our developed metadata schema from an Openlink's Virtuoso triple store, and a enhanced enrichment algorithm.

Via the enhanced metadata enrichment algorithm, the published records are enriched in two stages. First, the metadata is enriched with data inside the repository. This way, the metadata coming from different institutions in Flanders, Belgium, are interlinked. Secondly, the metadata records are enriched with data coming from the DBpedia and GeoNames datasets. By interlinking with these datasets, the metadata records are weaved into the web of data. This way, we interlink the data coming from the Flemish heritage institutions and publish these records as linked open data.

ACKNOWLEDGMENTS

The research activities that have been described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Flanders), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

REFERENCES

[1] Project Heritage2.0. Available at: http://projects.ibbt.be/erfgoed2.0

[2] C. Lagoze and H.V. de Sompel, The open archives initiative protocol for metadata harvesting – version 2.0, 2002. Available at: http://www.openarchives.org/OAI/openarchivesprotocol.html

[3] T. Berners-Lee, Linked Data, July 2006. Available at: http://www.w3.org/DesignIssues/LinkedData.html

[4] E. Prud'hommeaux and A. Seaborne, SPARQL Query Language for RDF, 2008. Available at: http://www.w3.org/TR/rdf-sparql-query/

[5] DBpedia. Available at: http://dbpedia.org/about

[6] GeoNames. Available at: http://www.geonames.org/export/

[7] B. Haslhofer and B. Schandl, The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data, Proceedings of the Linked Data on the Web Workshop, Beijing, China, 2008, CEUR Workshop Proceedings, ISSN 1613-0073, online CEUR-WS.org/Vol-369/

[8] Y. Lafon, HTTP - Hypertext Transfer Protocol, 2008. Available at: http://www.w3.org/Protocols/

[9] T. Bray, J. Paoli, C.M. Sperberg-McQueen, E. Maler, F. Yergeau, Extensible Markup Language (XML) 1.0 (fifth edition), 2008. Available at: http://www.w3.org/TR/REC-xml/

[10] Fedora Development Team, Fedora Open Source Repository Software: White Paper, 2005. Available at: http://www.fedora-commons.org/pdfs/WhitePaper.10.28.05.pdf

[11] The DSpace Foundation, DSpace 1.5.1 Manual, 2008. Available at: http://www.dspace.org/1_5_1Documentation/DSpace-Manual.pdf

[12] P. Millington and W.J. Nixon, EPrints 3 Pre-Launch Briefing, 2007. Available at: http://www.ariadne.ac.uk/issue50/eprints-v3-rpt/

[13] The European Library Project. Available at: http://www.theeuropeanlibrary.org

[14] R. Lasher and D. Cohen, A Format for Bibliographic Records, 1995. Available at: http://www.ietf.org/rfc/rfc1807.txt

[15] Library of Congress, MARC Standards homepage, 2008. Available at: http://www.loc.goc/marc

[16] Library of Congress, MARC 21 Format for Bibliographic Data, 2008. Available at: http://www.loc.gov/marc/bibliographic/

[17] Library of Congress, Metadata Object Description Schema, 2009. Available at: http://www.loc.gov/standards/mods/

[18] Library of Congress, Metadata encoding & Transmission Standard, 2009. Available at: http://www.loc.gov/standards/mets/

[19] M. Dean, D. Connolly, F. van Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider, L.A. Stein: OWL web ontology language reference. W3C Working Draft, 2003. Available at: http://www.w3.org/TR/2003/WD-owl-ref-20030331

[20] OAI2LOD Server, OAI2LOD Server homepage, 2008. Available at: http://www.mediaspaces.info/tools/oai2lod/

[21] C. Bizer and A. Seaborne, D2RQ – Treating non-RDF databases as virtual RDF Graphs, 2004. Available at: http://www.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

[22] D. Beckett and B. McBride, RDF/XML Syntax Specification, 2004. Available at: http://www.w3.org/TR/rdf-syntax-grammar/

[23] JENA – A Semantic Web Framework for Java. Available at: http://jena.sourceforge.net/index.html

[24] Openlink Software, Virtuoso open-source edition, 2009. Available at: http://virtuoso.openlinksw.com/wiki/main/Main

[25] D. Brickley, G.V. Guha, RDF Vocabulary Description Language 1.0: RDF Schema, 2004. Available at: http://www.w3.org/TR/rdf-schema/

Sam Coppens received his M.Sc. degree in Engineering from K.U. Leuven, Belgium, in 2005. He joined the Multimedia Lab in 2007 where he is researcher and involved in several cultural dissemination projects. His research interests include semantic web technologies, linked open data, and ontology design.

Erik Mannens received his Master's degree in engineering (1992) at KAHO Ghent and his Master's degree in computer science (1995) at K.U. Leuven University. His major expertise is centered on broadcasting, iDTV and web development. He is involved in several projects as senior researcher, he's co-chair of W3C's Media Fragments Working Group and actively participating in other W3C's semantic web standardization activities. He's also member of the technical committee of ACM MultiMedia, SAMT and MAReSO.

Rik Van de Walle received his M.Sc. and Ph.D. degrees in Engineering from Ghent University, Belgium in 1994 and 1998, respectively. After a visiting scholarship at the University of Arizona (Tucson, USA), he returned to Ghent University. In 2001 he became a professor at the Department of Electronics and Information Systems (Ghent University-IMEC, Belgium) and founded the Multimedia Lab. Rik Van de Walle has been/is editor of the following MPEG specifications: MPEG-21 Digital Item Declaration Language; MPEG-21 Digital Item Processing; MPEG-21 Digital Item Processing - Technologies under Consideration; and MPEG-21 Reference Software. Rik Van de Walle has been involved in the organization of and/or review of papers for several international conferences (e.g., IEEE ICME, WIAMIS, ISAS-SCI, ACIVS, Mirage, EUROMEDIA-Mediatec). His current research interests include multimedia content delivery, presentation and archiving, coding and description of multimedia data, content adaptation, interactive (mobile) multimedia applications, interactive digital TV. Multimedia Lab is one of the partners of the Interdisciplinary Institute for Broadband Technology (IBBT), which was founded by the Flemish Government in 2004.