ILastic: Linked Data Generation Workflow and User Interface for iMinds Scholarly Data

Anastasia Dimou[1,2](✉), Gerald Haesendonck[1], Martin Vanbrabant[1],
Laurens De Vocht[1], Ruben Verborgh[1], Steven Latré[2], and Erik Mannens[1]

[1] IDLab, IMEC, Ghent University, Ghent, Belgium
{anastasia.dimou,gerald.haesendonck,martin.vanbrabant,laurens.vocht,
ruben.verborgh,erik.mannens}@ugent.be
[2] IDLab, IMEC, University of Antwerp, Antwerpen, Belgium

Abstract. Enriching scholarly data with metadata enhances the publications' meaning. Unfortunately, different publishers of overlapping or complementary scholarly data neglect general-purpose solutions for metadata and instead use their own ad-hoc solutions. This leads to duplicate efforts and entails non-negligible implementation and maintenance costs. In this paper, we propose a reusable Linked Data publishing workflow that can be easily adjusted by different data owners to (i) generate and publish Linked Data, and (ii) align scholarly data repositories with enrichments over the publications' content. As a proof-of-concept, the proposed workflow was applied to the iMinds research institute data warehouse, which was aligned with publications' content derived from Ghent University's digital repository. Moreover, we developed a user interface to help lay users with the exploration of the iLastic Linked Data set. Our proposed approach relies on a general-purpose workflow. This way, we manage to reduce the development and maintenance costs and increase the quality of the resulting Linked Data.

1 Introduction

Semantic publishing (i) enhances the meaning of publications by enriching them with metadata, (ii) facilitates its automated discovery and summarization, (iii) enables its interlinking, (iv) provides access to data within the article in actionable form, and (v) facilitates its integration [20]. Scholarly publishing has undergone a digital revolution with massive uptake of online provision, but it has not realized the potential offered by the Web [20], let alone the Semantic Web. Even though the latter allows providing identifiers and machine-readable metadata for the so-called *enhanced-publications* [3], benefits come at a cost.

Ad-hoc solutions were established so far for generating Linked Data from scholarly data. Therefore, even though different data owners may hold overlapping or complementary data, new implementations are developed every time, customized to each publishers' infrastructure. Such approaches are adopted not

A. González-Beltrán et al. (Eds.): SAVE-SD 2017/2018, LNCS 10959, pp. 15–32, 2018.
https://doi.org/10.1007/978-3-030-01379-0_2

only by *individual efforts*, such as COLINDA[1] [21], but also by *publishing companies*, such as Springer[2] or the Semantic Web journal[3], as well as *large-scale initiatives*, such as OpenAIRE LOD[4] [23]. Nevertheless, this leads to duplicate efforts which entail non-negligible implementation and maintenance costs. The DBLP computer science bibliography[5] is one of the few exceptions, as it relies on an established approach, the D2RQ language [5] and its corresponding implementation, which is reusable and the Linked Data set is reproducible.

Workflows that semantically annotate scholarly data from repositories with structured data, generate Linked Data sets which remain independent, whereas the actual publications' content enrichment is rarely published as a Linked Data set. Besides the structured metadata regarding researchers and their publications, additional complementary information might be derived from the publications' actual content by extracting and semantically annotating it. While there are many approaches proposed for identifying entities in publications and associating them with well-known entities, such as [2,19] and others summarized at [11], publishing such metadata as a coherent Linked Data set and, even more, associating them with complementary metadata from repositories with structured data does not frequently and systematically happen so far. In this context, Bagnacani et al. [1] identified the most prevalent fragments of scholarly data publishing approaches: (i) bibliographic, (ii) authorship, and (iii) citations.

In this paper, we present a general-purpose Linked Data publishing workflow, adjusted to scholarly data publishing, which can be used by different data owners. The proposed workflow is applied in the case of the iLastic[6] Linked Data set generation and publishing for the iMinds[7] research institute's scholarly (meta)data. The workflow is complemented by an easily adjustable and extensible user interface which allows users to explore the underlying Linked Data. The scope is to align the Linked Data generation workflow for structured data with the plain text enrichment services developed in particular for iLastic.

The remainder of the paper is organized as follows: In the next section (Sect. 2), the state of the art is summarized. Then, in Sect. 3, the iLastic project is introduced, followed by the iLastic model (Sect. 4), the vocabularies used to annotate the data (Sect. 5) and the details about the generated Linked Data set (Sect. 6). Then the iLastic Linked Data generation workflow is presented (Sect. 7), followed by the iLastic user interface (Sect. 8). Last, in Sect. 9, we summarize our conclusions and our plans for future work.

[1] http://www.colinda.org/.
[2] http://lod.springer.com.
[3] http://semantic-web-journal.com/sejp/page/semanticWebJournal.
[4] http://beta.lod.openaire.eu/.
[5] http://dblp.l3s.de/dblp++.php.
[6] http://explore.ilastic.be.
[7] https://www.iminds.be/.

2 State of the Art

In this section, we indicatively mention a few existing solutions for scholarly data enrichment and its corresponding Linked Data set generation and publication.

COnference LInked DAta (See footnote 1) (COLINDA) [21] exposes information about scientific events, like conferences and workshops, for the period from 2007 up to 2013. It is one of the first Linked Data sets published on scholarly data, thus it is a custom solution which cannot be reused by any other data publisher who maintains similar data, as it might occur with the solution we propose. The data is derived from WikiCfP[8] and Eventseer[9]. COLINDA uses as input a harmonized and preprocessed CSV which contains data from the two aforementioned data sources. The CSV is imported into a MySQL database and a batch process is used to generate the corresponding Linked Data, whereas in our proposed solution, the CSV can be used directly and any data process may be applied during the Linked Data generation.

Even though OpenAIRE LOD [23], the Open Access Infrastructure for Research in Europe[10], was recently launched, it still relied on a custom solution to generate its Linked Data set from OpenAIRE Information Space which cannot be reused by any other data publisher. As performance and scalability were major concerns, a MapReduce [7] processing strategy was preferred. The original data is available in HBase[11], XML[12] and CSV[13] formats. Among the three formats, CSV was preferred to be used for the Linked Data set generation as it is not much slower than HBase, but it is much more maintenable [23]. Besides the CSV file which contains the actual data, an additional manually composed CSV is provided with relations about duplicate records.

The Semantic Lancet Project [1] publishes Linked Data for scholarly publications from Science Direct[14] and Scopus[15]. Its Linked Data set is generated relying on a series of custom scripts. Therefore, incorporating a new data source requires writing such a custom script, whereas in our solution, it is only required to provide the resource's description. Nevertheless, it is one of the few Linked Data sets for scholarly data whose Linked Data set is enhanced with more knowledge derived from the publications' content. This is achieved relying on FRED[16], a tool that parses natural language text, and implements deep machine learning.

DBLP computer science bibliography (DBLP) (See footnote 5) is one of the exceptions, as it relies on an established and, thus reusable and reproducible, approach to generate its Linked Data set. The FacetedDBLP (See footnote 5) is generated from data residing in DBLP databases by executing mapping rules

[8] http://www.wikicfp.com/cfp/.

[9] http://eventseer.net/.

[10] https://www.openaire.eu/.

[11] https://hbase.apache.org/.

[12] https://www.w3.org/TR/xml11/.

[13] https://tools.ietf.org/html/rfc4180.

[14] http://www.sciencedirect.com/.

[15] https://scopus.com/.

[16] http://wit.istc.cnr.it/stlab-tools/fred.

described in D2RQ mapping language [5], the predecessor of the W3C recommended R2RML [6], and published using a D2R server[17] instance. Nevertheless, D2RQ may only be used with data residing or imported in a database, whereas our solution may also support data in other structures derived from different access interfaces.

The Semantic Web Dog Food[18] (SWDF) contains metadata for the ESWC and ISWC Semantic Web conferences. Its Linked Data is generated from data derived from small size spreadsheets, tables or lists in documents, and HTML pages. The input data after being extracted, is turned into XML format which is further processed (i.e. cleansed) or non-RDF BibTeX and iCalendar documents. The former is produced manually using a generic XML editor and custom scripts were developed to generate the Linked Data. The latter allows to use some more automated tools, such as the bibtex2rdf converter[19] or Python scripts[20]. A detailed description process of the SWDF's generation is available at [18].

Lately, the SWDF dataset was migrated to Scholarly Data[21]. Conference Linked Open Data Generator[22] (cLODg) [14] is the tool used to generate the Scholarly Data Linked Data set. Besides, DBLP, this is one of the tools whose generated Linked Data set may be reproduced and the tool itself may be reused. It uses D2R conversions, as DBLP, but it also requires data derived from different data sources to be turned into CSV files which, on their turn, are ingested into a SQL database. With our proposed approach, we manage even to avoid this preprocessing step and directly use the original data sources [13], reducing the required effort and maintenance costs and increasing at the same time the reusability of our workflow and the reproducibility of the generated Linked Data.

3 The iLastic Project

The iLastic project was launched by the iMinds research institute in 2015 and aims to publish scholarly data which is associated with researchers affiliated with any of the iMinds labs. The iMinds labs are spread across Flanders' universities, thus researchers affiliated with iMinds are also affiliated with a university and their publications are archived by both iMinds and the corresponding university. To be more precise, iMinds maintains its own data warehouse (DWH) with metadata related to its researchers, the labs they belong to, publications they co-author, and projects they work on. The project aims to enrich information derived from data in the iMinds data warehouse with knowledge extracted from the publications' content. To achieve that, Flemish universities' digital repositories were considered, as they provide the full content of open accessed publications.

[17] http://d2rq.org/d2r-server.

[18] http://data.semanticweb.org/.

[19] http://www.l3s.de/~siberski/bibtex2rdf/.

[20] http://www.w3.org/2002/12/cal/.

[21] http://www.scholarlydata.org/.

[22] https://github.com/anuzzolese/cLODg2.

The project relies on (i) an in-house *general-purpose Linked Data generation workflow for structured data*, which was used for semantically annotating the data derived from the iMinds data warehouse; (ii) an in-house *publications retrieval and enrichment mechanism* developed for the project needs; and (iii) an *extensible and adjustable user interface* to facilitate non-Semantic Web expert users to search and explore the semantically enriched data.

The project was conducted in two phases:

Phase 1: Proof of Concept. In the first phase, the goal was to provide a *proof-of-concept* regarding the feasibility of the solution and its potential with respect to the expected results, namely showing the target milestones can be reached. In this phase, we mainly relied on selected data retrieved from the iMinds data warehouse, regarding persons, publications, organizations (universities and labs) and projects. Those entities formed the first version of the iLastic dataset.

Phase 2: Enrichment, Packaging and Automatization. In the second phase, two goals were posed: (i) enrich the first version of the iLastic Linked Data with knowledge extracted from the publications' content, and (ii) automate the Linked Data generation workflow to systematically generate Linked Data from the iMinds data warehouse, enrich them and publish them altogether. The complete workflow is now executed in the beginning of each month. In this phase, we packaged the solution, so other research institutes only need to configure their own rules for their data and repositories to generate their own Linked Data.

4 The iLastic Model

The iLastic dataset consists of data that describe (i) people, (ii) publications, (iii) projects and (iv) organizations. More details follow in this section about each type of entity, as well as challenges we had to deal with for the first two.

4.1 People

The iLastic dataset consists of data regarding people who work for iMinds, but not exclusively. They might be researchers, who belong to one of the iMinds labs and were authors of publications. Besides researchers affiliated with iMinds, many more people might appear in the iMinds data warehouse, even though they do not belong in any of the iMinds labs, thus they are not iMinds personnel but they co-authored one or more papers with one or more researchers from iMinds.

People are associated with their publications, their organizations, and, on rare occasions, with the projects they are involved in if their role is known, for instance if they are the project or research leads or the contact persons.

Challenges. iMinds personnel is identified with a unique identifier which the CRM system assigns to each person. However, researchers, who are co-authors in publications and do not belong to any of the iMinds labs, are not assigned such a unique identifier, as they are not iMinds personnel.

Therefore, there were three major challenges that we needed to address: (i) *distinguish iMinds researchers* from *non-iMinds researchers*; and (ii) among the non-iMinds researchers, *identify the same person* appearing in the dataset multiple times, being only aware of the researchers name (and on certain occasions their affiliation). Besides the data from the iMinds data warehouse, integrating information extracted from the papers' content required us to deal with one more challenge: (iii) *associate authors extracted from the publications' content* with the people that appear in the iMinds data warehouse.

4.2 Publications

The iLastic dataset also includes information regarding publications published by researchers when, at least one of the co-authors, is an iMinds researcher. As with iMinds researchers, each publication that is registered in iMinds data warehouse is assigned a unique identifier. Nevertheless, even though the iMinds data warehouse includes some information regarding publications, it refers mainly to metadata, such as the title, authors publication date or category. There is no information regarding the actual content of publications. To enrich the information regarding publications, we considered integrating data from complementary repositories, namely universities' repositories, such as Ghent University Academic Bibliography digital repository[23] or the digital repository for KU Leuven Association research[24]. These repositories also provide the PDF file of open access publications which can be parsed and analyzed to derive more information.

Challenges. There were two challenges encountered with respect to publications' semantic annotation: (i) *aligning* publications as they appear in the iMinds data warehouse with corresponding publications in universities' repositories, and (ii) *enriching* the structured data annotation derived from the iMinds data warehouse with plain text enrichment derived from the publications' actual content.

To be more precise, in the former case, we needed to define the proper algorithms and heuristics which allowed us to identify the publications' content by comparing the titles of the publications, as they appear in the iMinds data warehouse, with the titles as extracted from the publications' PDF. In the latter case, once the PDF of a certain publication was identified, the extraction of meaningful keywords, the recognition of well-known entities among those keywords, and the enrichment of the publications with this additional knowledge was required.

[23] https://biblio.ugent.be/.

[24] https://lirias.kuleuven.be/.

4.3 Organizations

The iMinds research institute is a multi-part organization which consists of several labs which are also associated with different universities in Flanders. The information about each one of the labs was required to be semantically annotated. Persons, publications and projects are linked to the different iMinds labs.

4.4 Projects

Last, a preliminary effort was put on semantically annotating the information related to projects the different iMinds labs are involved in. The projects are associated with people who work on them, but only the information regarding the projects' research and project leads, as well as contact person was considered.

5 The iLastic Vocabulary

We considered the following commonly used vocabularies to semantically annotate the iMinds scholarly data: BIBO[25], bibTex[26], CERIF[27], DC[28] and FOAF[29]. An indicative list of the high level classes used for the iLastic dataset is available at Table 1 and the most frequently used properties is available at Table 2.

The Bibliographic Ontology (BIBO) provides basic concepts and properties to describe citations and bibliographic references. The bibTeX ontology is used to describe bibTeX entries. The Common European Research Information Format (CERIF) ontology provides basic concepts and properties for describing research information as semantic data. The DCMI Metadata Terms (DC) includes metadata terms maintained by the Dublin Core Metadata Initiative to describe general purpose high level information. Last, the Friend Of A Friend (FOAF) ontology is used to describe people.

Table 1. Classes used to semantically annotate the main iLastic entities

cerif:Person
cerif:OrganizationalUnit
cerif:Publication
cerif:Project

Different vocabularies were used for different concepts. In particular, we used CERIF, DC and FOAF vocabularies to annotate data regarding people. The

[25] http://purl.org/ontology/bibo/.
[26] http://purl.org/net/nknouf/ns/bibtex.
[27] http://spi-fm.uca.es/neologism/cerif.
[28] http://dublincore.org/documents/dcmi-terms/.
[29] http://xmlns.com/foaf/0.1/.

Table 2. Properties used to semantically annotate the iLastic data model

Bibo/bibTeX	CERIF	DCTerms/FoaF/iM
bibo:identifier	cerif:internalidentifier	dcterms:identifier
bibo:abstract	cerif:linksToOrganisationUnit	dcterms:issued
bibo:issn	cerif:linksToPublication	foaf:familyName
bibo:isbn13	cerif:name	foaf:givenName
bibo:uri	cerif:title	im:webOfScience
bibtex:howPublished	cerif:acronym	im:publicationCategory

more generic DC and FOAF vocabularies were used to annotate information regarding, for instance, the name and surname of the author, whereas CERIF was used to define and associate with its organization and publications.

BIBO, BibTex, CERIF and DC vocabularies were used to annotate publications, FOAF, CERIF and DC to annotate organizational units and CERIF to annotate projects. Note, to cover cases where the aforementioned or other vocabularies did not have properties to annotate particular internal concepts of the iMinds data, we used custom properties defined for our case. For instance, iMinds tracks if a certain publication is indexed by Web Of Science[30]. Therefore, a custom property (`im:webOfScience`) was introduced to represent this knowledge. Moreover, iMinds classifies publications in different categories. A custom property (`im:publicationCategory`) was introduced for this purpose.

6 The iLastic Dataset

The iLastic dataset contains information about 59,462 entities. In particular, it contains information about 12,472 researchers (both people affiliated with iMinds and externals), 22,728 publications, 81 organizational units, and 3,295 projects. It consists of 765,603 triples in total and is available for querying at http://explore.ilastic.be/sparql.

7 The iLastic Linked Data Publishing Workflow

In this section, we describe the complete workflow for the generation of Linked Data sets from scholarly data, as it was applied in the case of iMinds.

The workflow consists of two pipelines: (i) one enriching the research metadata derived from the iMinds data warehouse, and (ii) one enriching the publications' content. The two pipelines aim to deal with the peculiarities of the different nature that the original data has, namely the structured data and plain text data, while they merge when the final Linked Data set is generated and published. An interface is built on top of the iLastic dataset offering a uniform

[30] http://www.webofknowledge.com/.

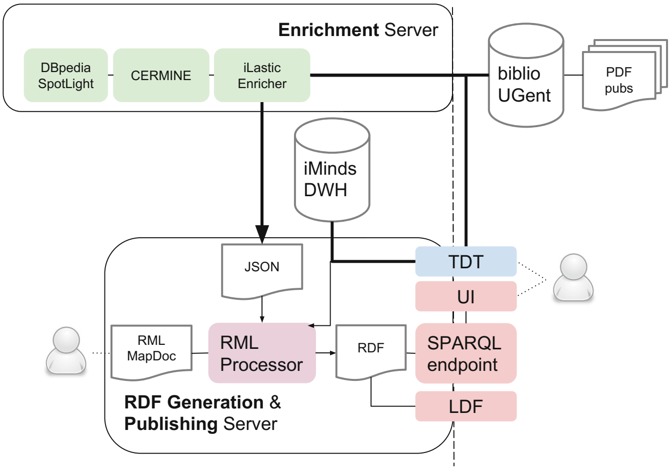


Fig. 1. Linked Data set generation and publishing workflow for the iLastic project.

interface to the users for searching and navigating within the iLastic dataset. The entire workflow consists of Open Source tools which are available for reuse.

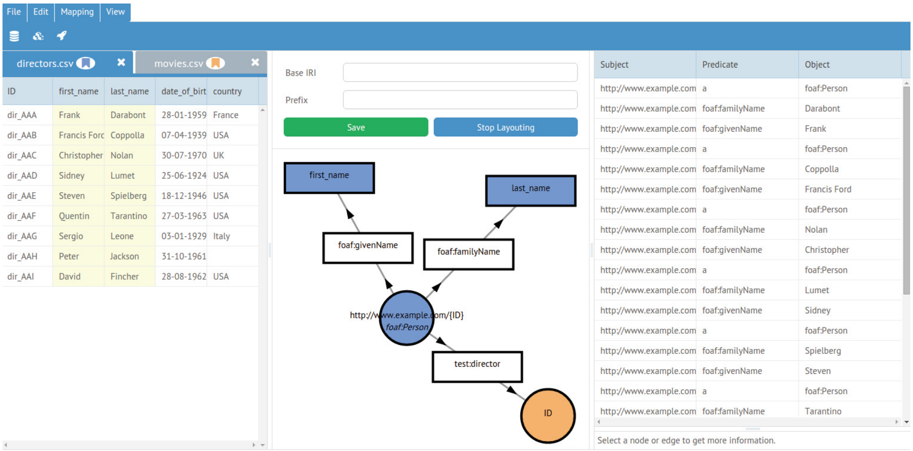
The Linked Data publication workflow for iLastic is presented at Fig. 1. Data is derived from the iMinds data warehouse via the DataTank. For each publication whose authors are affiliated with Ghent University, its corresponding one is identified in the Ghent University repository. Its PDF is then processed by the *iLastic Enricher* and its RDF triples are generated in combination with the information residing in the iMinds data warehouse. The data is published via a Virtuoso SPARQL endpoint and SPARQL templates are published via the DataTank which are used by the iLastic User Interface. Organizations which desire to adopt our proposed Linked Data generation and publication workflow may follow the corresponding tutorial [9]. Moreover, it is possible to extend the range of data sources depending on the use case. For instance, publications may be e-prints, or might be derived from an open repository.

The workflow is described in more details in the following subsections. In the next section, we explain how the rules to generate Linked Data are defined in the case of the iLastic Linked Data publishing workflow (Sect. 7.1). Then, we describe how data is retrieved, both from the iMinds data warehouse and the Ghent University digital repository in our exemplary use case (Sect. 7.2). The aforementioned input data and rules are used to generate the Linked Data, the iLastic Linked Data set in our use case, using our proposed workflow, as specified in Sect. 7.3, which is then published, as specified in Sect. 7.4, and accessed via a dedicated user interface, as described in Sect. 8. Last, the installation of our use case is briefly mentioned at Sect. 7.5.

7.1 Mapping Rules Definition

Generation. Firstly, we obtained a sample of the data derived from the iMinds data warehouse. We relied on this sample data to define the mapping rules that specify how the iLastic Linked Data is generated in our case. To facilitate the editing of mapping rules, we incorporate the RMLEditor [16]. If other organizations desire to reuse our proposed workflow, they only need to define their own mapping rules which refer to their own data sources. Defining such mapping rules for certain data, relying on target ontologies or existing mapping rules may be automated, e.g., as proposed by Heyvaert [15].

The RMLEditor[31] has a user friendly interface [17], as shown in Fig. 2, that supports lay users to define the mapping rules. The RMLEditor is used to generate the mapping documents for the data retrieved from the iMinds data warehouse. A *mapping document* summarizes the rules specifying how to generate the Linked Data. After all mapping rules are defined, we exported them from the RMLEditor. The RMLEditor exports the mapping rules expressed using the RDF mapping language (RML) [12] in a single mapping document.



ID	first_name	last_name	date_of_birt	country
dir_AAA	Frank	Darabont	28-01-1959	France
dir_AAB	Francis Forc	Coppolla	07-04-1939	USA
dir_AAC	Christopher	Nolan	30-07-1970	UK
dir_AAD	Sidney	Lumet	25-06-1924	USA
dir_AAE	Steven	Spielberg	18-12-1946	USA
dir_AAF	Quentin	Tarantino	27-03-1963	USA
dir_AAG	Sergio	Leone	03-01-1929	Italy
dir_AAH	Peter	Jackson	31-10-1961	
dir_AAI	David	Fincher	28-08-1962	USA

Subject	Predicate	Object
http://www.example.com	a	foaf:Person
http://www.example.com	foaf:familyName	Darabont
http://www.example.com	foaf:givenName	Frank
http://www.example.com	a	foaf:Person
http://www.example.com	foaf:familyName	Coppolla
http://www.example.com	foaf:givenName	Francis Ford
http://www.example.com	a	foaf:Person
http://www.example.com	foaf:familyName	Nolan
http://www.example.com	foaf:givenName	Christopher
http://www.example.com	a	foaf:Person
http://www.example.com	foaf:familyName	Lumet
http://www.example.com	foaf:givenName	Sidney
http://www.example.com	a	foaf:Person
http://www.example.com	foaf:familyName	Spielberg
http://www.example.com	foaf:givenName	Steven
http://www.example.com	a	foaf:Person
http://www.example.com	foaf:familyName	Tarantino

Select a node or edge to get more information.

Fig. 2. The RMLEditor user interface for editing rules that define how iLastic Linked Data set is generated.

Validation. The exported mapping document is validated for its consistency using the RMLValidator [10]. At this step, we make sure that the semantic annotations defined are consistent and no violations occur because multiple vocabularies are (re)used and combined. Any violations are addressed and the final mapping document is produced to be used for generating the Linked Data set.

[31] http://rml.io/RMLeditor.

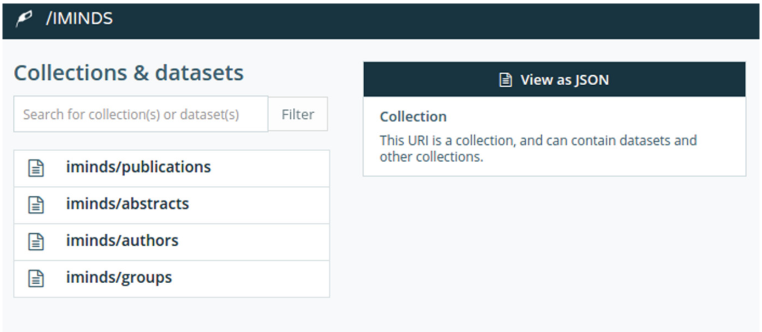
The mapping documents that were generated for the iLastic project are available at http://rml.io/data/iLastic/PubAuthGroup_Mapping.rml.ttl.

7.2 Data Access and Retrieval

The iLastic workflow consists of two input pipelines: (i) one for publishing structured data derived from the iMinds data warehouse, and (ii) one for publishing the results of the plain text enrichment. The two input pipelines are merged at the time of the Linked Data generation. Data originally residing at the iMinds data warehouse, as well as data derived from Ghent University digital repository, are considered to generate the iLastic Linked Data set.

Both input pipelines require accessing different parts of the data stored in the iMinds data warehouse. To achieve that, we published the corresponding SQL queries on a DataTank[32] instance that acts as the interface for accessing the underlying data for both pipelines. The DataTank offers a generic way to publish data sources and provides an HTTP API on top of them. The results of the SQL queries against the iMinds data warehouse and of the publications' enrichment are proxied by the DataTank and are returned in (paged) JSON format.

The original raw data as retrieved from the iMinds data warehouse and made available as Open Data can be found at http://explore.ilastic.be/iminds. The DataTank user interface is shown in Fig. 3.



This URI is a collection, and can contain datasets and other collections.

Fig. 3. The DataTank interface for accessing the iMinds data as raw Open Data.

7.3 Linked Data Generation

Structured Data Pipeline. The *structured-data pipeline* aims to semantically annotate data derived from the iMinds data warehouse. It considers the input data as it is retrieved from the DataTank and aims to directly

[32] http://thedatatank.com/.

semantically annotate them with the aforementioned vocabularies and ontologies. The RMLProcessor relies on machine-interpretable descriptions of the data sources [13]. To access the iMinds data warehouse, the RMLProcessor relies on the API's description which is defined using the Hydra vocabulary[33].

Plain Text Enrichment Pipeline. The *plain-text-enrichment pipeline* aims to enrich the publications metadata with information derived from the publications' actual content. Thus, retrieving, extracting and processing each publication's text is required. This occurs in coordination with the university repositories. To be more precise, for each publication, the university affiliated with the authors which is also part of iMinds is considered to retrieve the publication from its repository. For our exemplary case, the Ghent University API is considered[34].

For each publication that appears at the iMinds data warehouse and its authors are affiliated with a Ghent University lab, the corresponding publication is identified in the set of publications retrieved from the Ghent University API. The same publications that appear both in the iMinds data warehouse and Ghent University repository are identified applying fuzzy matching over their title and author(s), if the latter is available.

The fuzzy matching is performed in different successive steps. (i) Firstly normalisation is applied to the titles. For instance, punctuation and redundant white spaces are removed. (ii) Once the normalization is completed, exact matching based on string comparison is performed. (iii) If exact match fails, matching based on individual words is performed and the words position is also taken into account. For instance, matching 'Linked Data' and 'Linked Open Data' scores well, whereas 'Linked Data' and 'Data Linked' scores worse. (iv) If the score is below a threshold, another matching algorithm is performed to avoid mismatching due to typos. The latter eliminates the same words from both titles (these get a high score) and compare the remaining words on a character basis. Dealing with typos, e.g., 'Lined' instead of 'Linked', acronyms, e.g., 'DQ' for 'Data Quality', and prefixes, such as 'Special issue on ... : <title>', were the most challenging cases we addressed.

As soon as a publication is retrieved, it is assessed whether it is required to be processed or not. It is checked if its PDF is open accessed and if it is, then it is checked if it is an old publication and it was already processed based on the last modified date of the PDF file. If it is open accessed and not processed before, the PDF is retrieved for further processing. Information extracted from the PDF, such as keywords or authors, may also be used to enrich the information derived from the data warehouse if such data is missing or is not complete.

The *iLastic Enricher* consists of two main components: CERMINE and DBpedia spotlight. Those two tools were chosen based on the 2015 Semantic Publishing Challenge results [8]. To be more precise, the former was the challenge's best performing tool, while the latter was broadly used by several solutions every year the challenge was organized [11] In the case of the iLastic Linked

[33] http://www.hydra-cg.com/spec/latest/core/.

[34] https://biblio.ugent.be/doc/api.

Data generation and publication workflow, each retrieved PDF file is fed to the *iLastic Enricher*. The *iLastic Enricher* uses the Content ExtRactor and MINEr (CERMINE[35]) [22] to extract the content of the corresponding PDF file. As soon as the publication's content is retrieved, its abstract and main body are fed to DBpedia spotlight[36] to identify and annotate entities that also appear in the DBpedia dataset. Besides the abstract and main body of the publication, the keywords assigned by the users are also extracted and annotated by DBpedia spotlight.

The output is summarized in a JSON file, where all identified terms are summarized. Such JSON files may be found at http://rml.io/data/iLastic/. The JSON file is passed to the RMLProcessor together with the rest of the data retrieved from the iMinds data warehouse, and the mapping document defined using the RMLEditor, to generate the resulting triples. This way, the corresponding publications information is enriched with data from its own text. Moreover, in cases where data derived from the iMinds data warehouse is missing, e.g., authors, the information is extracted from the publications. This way, not only the iLastic Linked Data set is enriched, but its completeness is also improved.

7.4 Linked Data Publication

Once the iLastic Linked Data set is generated, it is stored and published to a Virtuoso instance[37] which is installed on the same server for this purpose. Virtuoso is a cross-platform server that provides a triplestore and a SPARQL endpoint for querying the underlying Linked Data. This endpoint may be used by clients which desire to access the iLastic dataset, as it is used by the DataTank to provide data to the iLastic user interface –described in the next section (Sect. 8).

7.5 Current Installation

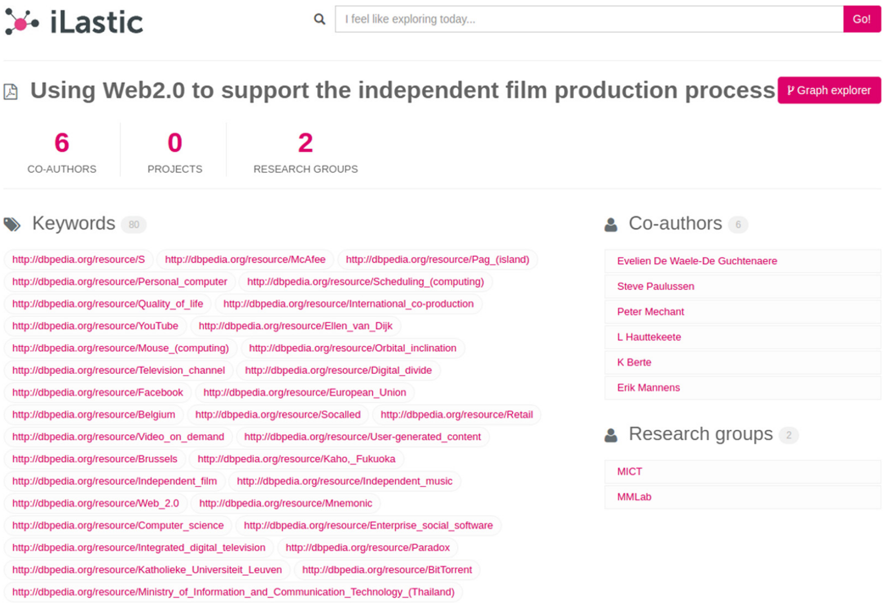
The iLastic Linked Data generation and publishing workflow consists of CERMINE which is available at https://github.com/CeON/CERMINE and DBpedia sportlight which is available at https://github.com/dbpedia-spotlight/dbpedia-spotlight for PDF extraction and annotation; the RMLProcessor which is available at https://github.com/RMLio/RML-Processor and RMLValidator which is available at https://github.com/RMLio/RML-Validator for the structured data annotation and alignment with non-structured data annotations; and the virtuoso endpoint which is available at https://github.com/openlink/virtuoso-opensource and DataTank which is available at https://github.com/tdt/ for data publishing.

The iLastic Linked Data publishing workflow runs on two servers. One accommodates the main part of the workflow. The data extraction and Linked Data

[35] http://cermine.ceon.pl/.
[36] https://github.com/dbpedia-spotlight/dbpedia-spotlight.
[37] https://github.com/openlink/virtuoso-opensource.

generation occurs there, namely the RMLProcessor runs there, as well as the publishing infrastructure are installed there, namely the Virtuoso instance, and the user interface. It runs on Ubuntu 14.04, with PHP 5.5.19, Java 1.7, MySQL 5.5, Virtuoso 7.20 and Nginx. Note the RMLEditor is used as a service residing on a different server, as it may be reused by other data owners too.

The publications enrichment, namely the *iLastic Enricher*, only takes place on a separate server, due to higher memory requirements. The server runs Debian GNU/Linux 7 with DBpedia Spotlight 0.7 and CERMINE installed.



Using Web2.0 to support the independent film production process

6 CO-AUTHORS

0 PROJECTS

2 RESEARCH GROUPS

Keywords 80

http://dbpedia.org/resource/S
http://dbpedia.org/resource/McAfee
http://dbpedia.org/resource/Pag_(island)
http://dbpedia.org/resource/Personal_computer
http://dbpedia.org/resource/Scheduling_(computing)
http://dbpedia.org/resource/Quality_of_life
http://dbpedia.org/resource/International_co-production
http://dbpedia.org/resource/YouTube
http://dbpedia.org/resource/Ellen_van_Dijk
http://dbpedia.org/resource/Mouse_(computing)
http://dbpedia.org/resource/Orbital_inclination
http://dbpedia.org/resource/Television_channel
http://dbpedia.org/resource/Digital_divide
http://dbpedia.org/resource/Facebook
http://dbpedia.org/resource/European_Union
http://dbpedia.org/resource/Belgium
http://dbpedia.org/resource/Socalled
http://dbpedia.org/resource/Retail
http://dbpedia.org/resource/Video_on_demand
http://dbpedia.org/resource/User-generated_content
http://dbpedia.org/resource/Brussels
http://dbpedia.org/resource/Kaho,_Fukuoka
http://dbpedia.org/resource/Independent_film
http://dbpedia.org/resource/Independent_music
http://dbpedia.org/resource/Web_2.0
http://dbpedia.org/resource/Mnemonic
http://dbpedia.org/resource/Computer_science
http://dbpedia.org/resource/Enterprise_social_software
http://dbpedia.org/resource/Integrated_digital_television
http://dbpedia.org/resource/Paradox
http://dbpedia.org/resource/Katholieke_Universiteit_Leuven
http://dbpedia.org/resource/BitTorrent
http://dbpedia.org/resource/Ministry_of_Information_and_Communication_Technology_(Thailand)

Co-authors 6

Evelien De Waele-De Guchtenaere
Steve Paulussen
Peter Mechant
L Hauttekeete
K Berte
Erik Mannens

Research groups 2

MICT
MMLab

Fig. 4. A publication as presented at iLastic user interface with information derived both from iMinds data warehouse and the analyzed and enriched publication's content.

8 The iLastic User Interface

The iLastic user interface was included in the second phase of the project aiming to make the iLastic dataset accessible to non-Semantic Web experts who do not have the knowledge to query it via its endpoint and to showcase its potential.

Users of the iLastic interface may discover knowledge resulting of the combination of the two channels of information. Users may search for iMinds researchers, discover the group they belong to, the publications they co-authored, the research areas they are active in, other people they collaborate with and, thus, their network of collaborators. Moreover, users may look for publications and discover combined information derived from the publications metadata derived from the iMinds data warehouse, such are the publication's category, as well as the keywords and main entities derived from the publication's content.

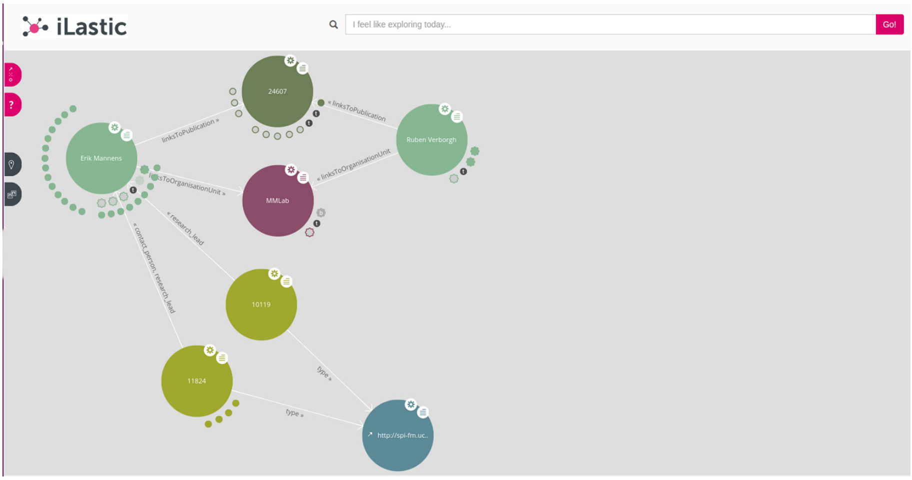


Fig. 5. The graph explorer view for Erik Mannens.

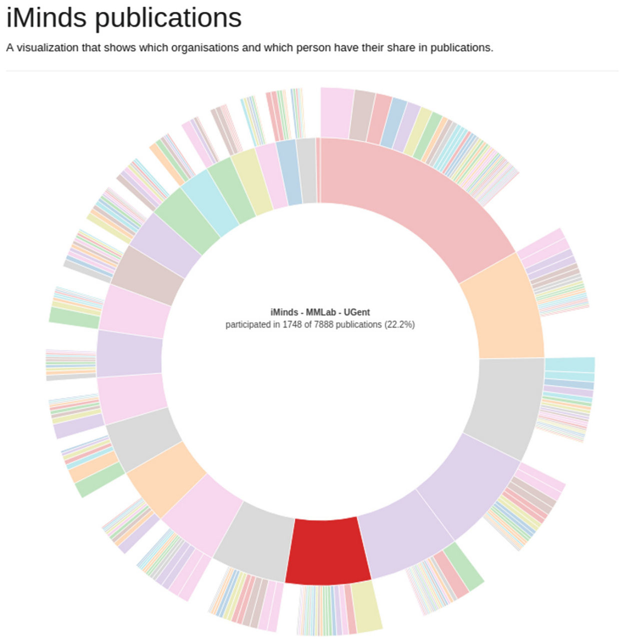
The iLastic user interface allows users to explore the different entities either using its regular interface or the graph explorer, or search within the Linked Data set. While the users explore the dataset via the user interface, their requests are translated in SPARQL queries which, on their own turn, are parameterized and published at the DataTank. Moreover, the search is supported by the iLastic sitemap. Both of them are explained in more details below.

The iLastic user interface relies on LodLive[38] [4], a demonstration of Linked Data standards' use to browse different resources of a dataset. The iLastic user interface can be accessed at http://explore.iLastic.be and a sreencast showcasing its functionality is available at https://youtu.be/ZxGrHnOuSvw.

Users may search for data in the iLastic Linked Data set. The iLastic sitemap was incorporated to support searching. It has a tree-like structure including the different entities handled in the iLastic project. This tree structure is indexed and serves as a search API, whose results are then used by the user interface's search application. The iLastic search application builds a front-end around the search API results where users can search for a person, publication or organization.

Moreover, a user may access the iLastic user interface to explore the integrated information on publications, as shown in Fig. 4. Besides the regular user interface, the users may take advantage of the incorporated graph explorer. For each one of the iLastic Linked Data set's entities, the user may switch from the regular interface to the graph explorer and vice versa. For instance, the graph explorer for 'Erik Mannens' is shown in Fig. 5. Last, a user may not only search for different entities within the iLastic Linked Data set, but also some preliminary analysis of the dataset's content is visualized, as shown in Fig. 6.

[38] http://en.lodlive.it/.



iMinds publications

A visualization that shows which organisations and which person have their share in publications.

iMinds - MMLab - UGent

participated in 1748 of 7888 publications (22.2%)

Fig. 6. The analyisis of iLastic Linked Data set.

9 Conclusions and Future Work

In this paper, we show how a general-purpose Linked Data generation workflow is adjusted to also generate Linked Data from raw scholarly data. Relying on such general-purpose workflows allows different data owners of scholarly data to reuse the same installations and re-purpose existing mapping rules to their own needs. This way, the implementation and maintenance costs are reduced.

In the future, we plan to extend the dataset with more data derived from both the iMinds research institute and the publications, such as references.

Acknowledgments. The described research activities were funded by Ghent University, imec, Flanders Innovation and Entrepreneurship (AIO), the Research Foundation – Flanders (FWO), and the European Union (EU).

References

1. Bagnacani, A., Ciancarini, P., Di Iorio, A., Nuzzolese, A.G., Peroni, S., Vitali, F.: The semantic lancet project: a linked open dataset for scholarly publishing. In: Lambrix, P. (ed.) EKAW 2014. LNCS (LNAI), vol. 8982, pp. 101–105. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-17966-7_10
2. Bahar, S., Witte, R.: Semantic representation of scientific literature: bringing claims, contributions and named entities onto the linked open data cloud. PeerJ Comput. Sci. **1**, e37 (2015)

3. Bardi, A., Manghi, P.: Enhanced publication management systems: a systemic approach towards modern scientific communication. In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015 Companion, pp. 1051–1052. ACM (2015)
4. Camarda, D.V., Mazzini, S., Antonuccio, A.: LodLive, exploring the web of data. In: Proceedings of the 8th International Conference on Semantic Systems, I-SEMANTICS 2012, pp. 197–200. ACM (2012)
5. Cyganiak, R., Bizer, C., Garbers, J., Maresch, O., Becker, C.: The D2RQ mapping language. Technical report, March 2012. http://d2rq.org/d2rq-language
6. Das, S., Sundara, S., Cyganiak, R.: R2RML: RDB to RDF mapping language. W3C Recommendation, W3C (2012). http://www.w3.org/TR/r2rml/
7. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
8. Iorio, A.D., Lange, C., Dimou, A., Vahdati, S.: Semantic publishing challenge – assessing the quality of scientific output by information extraction and interlinking. In: Gandon, F., Cabrio, E., Stankovic, M., Zimmermann, A. (eds.) SemWebEval 2015. CCIS, vol. 548, pp. 65–80. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25518-7_6
9. Dimou, A., Heyvaert, P., Taelman, R., Verborgh, R.: Modeling, generating, and publishing knowledge as linked data. In: Ciancarini, P., et al. (eds.) EKAW 2016. LNCS (LNAI), vol. 10180, pp. 3–14. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58694-6_1
10. Dimou, A., et al.: Assessing and refining mappingsto RDF to improve dataset quality. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS, vol. 9367, pp. 133–149. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25010-6_8
11. Dimou, A., Vahdati, S., Di Iorio, A., Lange, C., Verborgh, R., Mannens, E.: Challenges as enablers for high quality linked data: insights from the semantic publishing challenge. PeerJ Comput. Sci. **3**, e105 (2017)
12. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: a generic language for integrated RDF mappings of heterogeneous data. In: Proceedings of the 7th Workshop on Linked Data on the Web, CEUR Workshop Proceedings, vol. 1184 (2014)
13. Dimou, A., Verborgh, R., Sande, M.V., Mannens, E., Van de Walle, R.: Machine-interpretable dataset and service descriptions for heterogeneous data access and retrieval. In: Proceedings of the 11th International Conference on Semantic Systems, SEMANTICS 2015, pp. 145–152. ACM (2015)
14. Gentile, A.L., Nuzzolese, A.G.: cLODg - conference linked open data generator. In: Villata, S., Pan, J., Dragoni, M. (eds.) International Semantic Web Conference (Posters and Demos), CEUR Workshop Proceedings, vol. 1486. CEUR-WS.org (2015)
15. Heyvaert, P., Dimou, A., Verborgh, R., Mannens, E.: Ontology-based data access mapping generation using data, schema, query, and mapping knowledge. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) ESWC 2017. LNCS, vol. 10250, pp. 205–215. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58451-5_15
16. Heyvaert, P., et al.: RMLEditor: a graph-based mapping editor for linked data mappings. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9678, pp. 709–723. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-34129-3_43

17. Heyvaert, P., Dimou, A., Verborgh, R., Mannens, E., Van de Walle, R.: Towards a uniform user interface for editing mapping definitions. In: Workshop on Intelligent Exploration of Semantic Data (2015)
18. Möller, K., Heath, T., Handschuh, S., Domingue, J.: Recipes for semantic web dog food—the ESWC and ISWC metadata projects. In: Aberer, K. (ed.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 802–815. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_58
19. Ronzano, F., Saggion, H.: Dr. Inventor framework: extracting structured information from scientific publications. In: Japkowicz, N., Matwin, S. (eds.) DS 2015. LNCS (LNAI), vol. 9356, pp. 209–220. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24282-8_18
20. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learn. Publ. **22**(2), 85–94 (2009)
21. Softic, S., De Vocht, L., Mannens, E., Ebner, M., Van de Walle, R.: COLINDA: modeling, representing and using scientific events in the web of data. In: Proceedings of the 4th International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2015), CEUR Workshop Proceedings, vol. 1363 (2015)
22. Tkaczyk, D., Bolikowski, Ł.: Extracting contextual information from scientific literature using CERMINE system. In: Gandon, F., Cabrio, E., Stankovic, M., Zimmermann, A. (eds.) SemWebEval 2015. CCIS, vol. 548, pp. 93–104. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25518-7_8
23. Vahdati, S., Karim, F., Huang, J.-Y., Lange, C.: Mapping large scale research metadata to linked data: a performance comparison of HBase, CSV and XML. In: Garoufallou, E., Hartley, R.J., Gaitanou, P. (eds.) MTSR 2015. CCIS, vol. 544, pp. 261–273. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24129-6_23