

Mapping Large Scale Research Metadata to Linked Data: A Performance Comparison of HBase, CSV and XML

Sahar Vahdati¹, Farah Karim¹, Jyun-Yao Huang², and Christoph Lange³

¹ University of Bonn vahdati@uni-bonn.de, Karim@iai.uni-bonn.de

² National Chung Hsing University, Taiwan allen501pc@gmail.com

³ University of Bonn & Fraunhofer IAIS, Germany math.semantic.web@gmail.com

Abstract OpenAIRE, the Open Access Infrastructure for Research in Europe, comprises a database of all EC FP7 and H2020 funded research projects, including metadata of their results (publications and datasets). These data are stored in an HBase NoSQL database, post-processed, and exposed as HTML for human consumption, and as XML through a web service interface. As an intermediate format to facilitate statistical computations, CSV is generated internally. To interlink the OpenAIRE data with related data on the Web, we aim at exporting them as Linked Open Data (LOD). The LOD export is required to integrate into the overall data processing workflow, where derived data are regenerated from the base data every day. We thus faced the challenge of identifying the best-performing conversion approach. We evaluated the performances of creating LOD by a MapReduce job on top of HBase, by mapping the intermediate CSV files, and by mapping the XML output.

1 Introduction

The European Commission emphasizes open access as a key tool to bring together people and ideas in a way that catalyses science and innovation. More than ever before, there is a recognized need for digital research infrastructures for all kinds of research outputs, across disciplines and countries. OpenAIRE, the Open Access Infrastructure for Research in Europe (<http://www.openaire.eu>), (1) manages scientific publications and associated scientific material via repository networks, (2) aggregates Open Access publications and links them to research data and funding bodies, and (3) supports the Open Access principles via national helpdesks and comprehensive guidelines.

Data related to those in the OpenAIRE information space exist in different places on the Web. Combining them with OpenAIRE will enable new use cases. For example, understanding changes of research communities or the emergence of scientific topics not only requires metadata about publications and projects, as provided by OpenAIRE, but also data about events such as conferences as well as a knowledge model of research topics and subjects (cf. [1]).

The availability of data that is free to use, reuse and redistribute (i.e. *open data*) is the first prerequisite for analysing such information networks. However,

the diverse data formats and means to access or query data, the use of duplicate identifiers, and the heterogeneity of metadata schemas pose practical limitations on reuse. Linked Data, based on the RDF graph data model, is now increasingly accepted as a lingua franca to overcome such barriers [2].

The University of Bonn is coordinating the effort of publishing the OpenAIRE data as Linked Open Data (LOD) and linking it to related datasets in the rapidly growing LOD Cloud⁴. This effort is further supported by the Athena Research and Innovation Center and CNR-ISTI. Besides data about scientific events and subject classification schemes, relevant data sources include public sector information (e.g., to find research results based on the latest employment statistics, or to answer questions such as ‘how do the EU member states’ expenses for health research compare to their health care spendings?’) and open educational resources (‘how soon do emergent research topics gain wide coverage in higher education?’).

Concrete steps towards this vision are (1) mapping the OpenAIRE data model to suitable standard LOD vocabularies, (2) exporting the objects in the OpenAIRE information space as a LOD graph and (3) facilitating integration with related LOD graphs. Expected benefits include

- enabling semantic search over the outputs of European research projects,
- simplifying the way the OpenAIRE data can be enriched by third-party services, and consumed by interested data or service providers,
- facilitated outreach to related open content and open data initiatives, and
- enriching the OpenAIRE information space itself by exploiting how third parties will use its LOD graph.

The specifically tailored nature of the OpenAIRE infrastructure, its large amount of data (covering more than 11 million publications) and the frequent updates of the more than 5000 repositories from which the data is harvested pose high requirements on the technology chosen for mapping the OpenAIRE data to LOD. We therefore compared in depth three alternative mapping methods, one for each source format in which the data are available: HBase, CSV and XML.

Section 2 introduces the OpenAIRE data model and the three existing data sources. Section 3 presents our specification of the OpenAIRE data model as an RDF vocabulary. Section 4 establishes requirements for the mapping. Section 5 presents the state of the art for each of the three mapping approaches. Section 6 explains our three implementations. In section 7 we evaluate them in comparison, with regard to different metrics induced by the requirements. Section 8 reviews work related to our overall approach (comparing mappings and producing research LOD). Section 9 concludes and outlines future work.

2 Input Data

The data model of OpenAIRE infrastructure is specified as an entity relationship model (ERM) [3,4] with the following entity categories:

⁴ <http://lod-cloud.net>

- **Main entities** (cf. figure 1)⁵: Result (Publication or Dataset), Person, Organization, Projects, and DataSource (e.g. Repository, Dataset Archive or CRIS⁶). Instances of these are continuously harvested from data providers.
- **Structural entities** representing complex information about main entities: Instances (of a Result in different DataSources), WebResources, Titles, Dates, Identities, and Subjects.
- **Static entities**, whose metadata do not change over time: Funding. E.g., once a funding agency has opened a funding stream, it remains static.
- **Linking entities** represent relationships between entities that carry further metadata; e.g., an entity of type Person_Result whose property *ranking* has the value 1 indicates the first author.

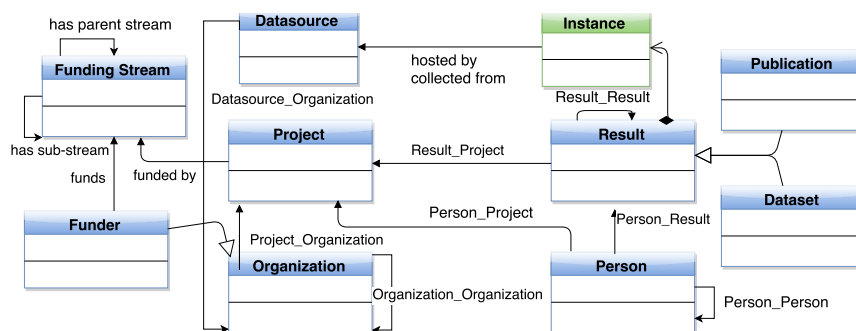


Figure 1: OpenAIRE Data Model: core entities and relationships

So far, the OpenAIRE data have been available in three formats: HBase, CSV and XML.

2.1 HBase

Currently, the master source of all OpenAIRE data is kept in HBase, a column store based on HDFS (Hadoop Distributed File System). HBase was introduced in 2012 when data integration efforts pushed the original PostgreSQL database to its limits: joins became inefficient and parallel processing, as required for de-duplication, was not supported. Each row of the HBase table has a unique row key and stores a main entity and a number of related linked entities. The attribute values of the main entities are stored in the *<family>:body* column, where the *<family>* is named after the type of the main entity, e.g., *result*, *person*,

⁵ https://issue.openaire.research-infrastructures.eu/projects/openaire2020-wiki/wiki/Core_Data_Model

⁶ Current research information system, a system to manage information about the research activity of an institution

project, *organization* or *datasource*. The attribute values of linked entities, indicating the relationship between main entities, are stored in dedicated column families $\langle family \rangle : \langle column \rangle$, where $\langle family \rangle$ is the class of the linked entity and $\langle column \rangle$ is the row key of the target entity. Both directions of a link are represented. Cell values are serialized as byte arrays according to the Protocol Buffers [5] specification; for example:

```
message Person {
  optional Metadata metadata = 2;
  message Metadata {
    optional StringField firstname = 1;
    repeated StringField secondnames = 2;
    optional Qualifier nationality = 9; ... }
  repeated Person coauthors = 4; }
```

The following table shows a publication and its authors. For readability, we abbreviated row keys and spelled out key-value pairs rather than showing their binary serialization.

RowKey	result: body	person: body	...hasAuthor:		...isAuthorOf:
50 ...0 01::39 b9...	resulttype= "publication"; title="The Data Model of ..."; dateofacceptance= "2012-01-01"; language="en"; publicationDate= "2012"; publisher= "Springer";		30 ...001::9897...	30 ...001::ef29...	50 ...001::39b9...
			ranking=1;	ranking=2;	
30 ...0 01::98 97...		firstname="Paolo"; lastname="Manghi";			ranking=1;
30 ...0 01::ef 29...		firstname="Nikos"; lastname="Houssos";			ranking=2;

2.2 CSV

CSV files aid the computation of statistics on the OpenAIRE information space. HBase is a sparse key value-store designed for data with little or no internal relations. Therefore, it is impossible to run complex queries directly on top of HBase, for example a query to find all results of a given project. It is thus necessary to transform the data to a relational representation, which is comprehensible for statistics tools and enables effective querying. Via an intermediate CSV representation, the data is imported into a relational database, which is queried for computing the statistics.

In this generation process, each main entity type (result, project, person, organization, datasource) is mapped to a CSV file of the same name, which is later imported into a relational database table. Each single-valued attribute of an entity (id, title, publication year, etc.) becomes a field in the entity's table. Multi-valued attributes, such as the publication languages of a result, are mapped to relation tables (e.g. `result_languages`) that represent a one-to-many relation between entity and attributes. Linked entities, e.g. the authors of a *result*, are represented similarly. As the data itself includes many special characters, for example commas in publication titles, the OpenAIRE CSV files use ! as a delimiter and wrap cell values into leading and trailing hashes:

```
#dedup_wf_001::39b91277f9a2c25b1655436ab996a76b#!#The Data Model of the OpenAIRE
Scientific Communication e-Infrastructure#!#null#!#null#!#Springer#!#null#!#null
#!#null#!#null#!#2012#!#2012-01-01#!#Open Access#!#Open Access#!#Access#!#null#!#
0#!#null#!#null#oai:http://helios-eie.ekt.gr:#!#publication#10442/13187oai:pumaoai.
isti.cnr.it:cnr.isti/cnr.isti/2012-A2-040#!#1#!
```

Finally, using CSV has the advantage that existing tools such as Sqoop can be used, thus reducing the need to develop and maintain customly implemented components on the OpenAIRE production system.

2.3 XML

OpenAIRE features a set of HTTP APIs⁷ for exporting metadata as XML for easy reuse by web services. These APIs use an XML Schema implementation of the OpenAIRE data model called OAF (OpenAIRE Format)⁸, where each record represents one entity. There is one API for searching, and one for bulk access. For example, the listing below comes from http://api.openaire.eu/search/publications?openairePublicationID=dedup_wf_001::39b91277f9a2c25b1655436ab996a76b and shows the metadata of a publication that has been searched for.

```
<oaf:result>
  <title schemename="dnet:dataCite_title" classname="main title"
    schemeid="dnet:dataCite_title" classid="main title">The Data Model of the
    OpenAIRE Scientific Communication e-Infrastructure</title>
  <dateofacceptance>2012-01-01</dateofacceptance>
  <publisher>Springer</publisher>
  <resulttype schemename="dnet:result_typologies" classname="publication"
    schemeid="dnet:result_typologies" classid="publication"/>
  <language schemename="dnet:languages" classname="English"
    schemeid="dnet:languages" classid="eng"/>
  <format>application/pdf</format>
  ...
</oaf:result>
```

The API for bulk access uses OAI-PMH (The **O**pen **A**rchives **I**nitiative **P**rotocol for **M**etadata **H**arvesting)⁹ to publish metadata and its corresponding endpoint is at http://api.openaire.eu/oai_pmh. The bulk access API lets developers fetch the whole XML files step by step. For our experiments, we obtained the XML data directly from the OpenAIRE server, as an uncompressed Hadoop Sequence-File¹⁰ comprising 500 splits of ~300 MB each.

3 Implementing the OpenAIRE Data Model in RDF

As the schema of the OpenAIRE LOD we specified an RDF vocabulary by mapping the entities of the ER data model to RDF classes and its attributes

⁷ <http://api.openaire.eu/>

⁸ <https://www.openaire.eu/schema/0.2/doc/oaf-0.2.html>

⁹ <http://www.openarchives.org/OAI/openarchivesprotocol.html>

¹⁰ <http://wiki.apache.org/hadoop/SequenceFile>

and relationships to RDF properties. We reused suitable existing RDF vocabularies identified by consulting the Linked Open Vocabularies search service¹¹ and studying their specifications. Reused vocabularies include Dublin Core for general metadata, SKOS¹² for classification schemes and CERIF¹³ for research organizations and activities. We linked new, OpenAIRE-specific terms to reused ones, e.g., by declaring *Result* a superclass of <http://purl.org/ontology/bibo/Publication> and <http://www.w3.org/ns/dcat#Dataset>.

We keep the URIs of the LOD resources (i.e. entities) in the <http://lod.openaire.eu/data/> namespace. We modelled them after the HBase row keys. In OpenAIRE, these are fixed length identifiers of the form $\{typePrefix\}|\{namespacePrefix\}::md5hash$. *typePrefix* is a two digit code, 10, 20, 30, 40 or 50, corresponding to the main entity types *datasource*, *organization*, *person*, *project* and *result*. The *namespacePrefix* is a unique 12-character identifier of the data source of the entity. For each row, *md5hash* is computed from the entity attributes. The resulting URIs look like http://lod.openaire.eu/data/result/dedup_wf_001::39b91277f9a2c25b1655436ab996a76b.

The following listing shows our running example in RDF/Turtle syntax.

```
@prefix oad: <http://lod.openaire.eu/data/> .
@prefix oav: <http://lod.openaire.eu/vocab#> .
# further prefixes omitted; see http://prefix.cc for their standard bindings.

oad:result/...001::39b9... rdf:type oav:Result, bibo:Publication;
  dct:terms:title "The Data Model of the OpenAIRE Scientific Communication
    e-Infrastructure"@en ;
  dct:terms:dateAccepted "2012-01-01"^^xsd:date ;
  dct:terms:language "en";
  oav:publicationYear 2012 ;
  dct:terms:publisher "Springer";
  dct:terms:creator oad:person/...001::9897..., oad:person/...001::ef29... .
oad:person/...001::9897... rdf:type foaf:Person;
  foaf:firstName "Paolo"; foaf:lastName "Manghi";
  oav:isAuthorOf oad:result/...001::39b9... .
oad:person/...001::ef29... rdf:type foaf:Person;
  foaf:firstName "Nikos"; foaf:lastName "Houssos";
  oav:isAuthorOf oad:result/...001::39b9... .
```

4 Requirements

In cooperation with the other technical partners in the OpenAIRE2020 consortium, most of whom had been working on the infrastructure in previous projects for years, we established the following requirements for the LOD export:

¹¹ <http://lov.okfn.org>

¹² <http://www.w3.org/2004/02/skos/>

¹³ Common European Research Information Format; see <http://www.eurocris.org/cerif/main-features-cerif>

- R1 The LOD output must follow the vocabulary specified in section 3.
- R2 The LOD must be generated from one of the three existing data sources, to avoid extra pre-processing costs.
- R3 The mapping to LOD should be maintainable w.r.t. planned extensions of the OpenAIRE data model (such as linking publications and data to software) and the evolution of linked data vocabularies.
- R4 The mapping to LOD should be orchestrable together with the other existing OpenAIRE data provision workflows, always exposing a consistent view on the information space, regardless of the format.
- R5 To enable automatic and manual checks of the consistency and correctness of the LOD before its actual publication, it should be made available in reasonable time in a private space.

To prepare an informed decision on the preferred input format to use for the LOD export, we realised one implementation for each of HBase, CSV and XML.

5 Technical State of the Art

For each possible approach, i.e. mapping HBase, CSV or XML to RDF, we briefly review the state of the art to give an overview of technology we could potentially reuse or build on, whereas section 8 reviews work related to our overall approach. We assess reusability w.r.t. the OpenAIRE-specific requirements stated above.

HBase, being a sparse, distributed and multidimensional persistent sorted map, provides dynamic control over the data format and layout. Several works have therefore explored the suitability of HBase as a triple store for semi-structured and sparse RDF data. Sun et al. adopted the idea of the Hexastore indexing technique for storing RDF in HBase [6]. Khadilkar et al. focused on a distributed RDF storage framework based on HBase and Jena to gain scalability [7]. Others have provided MapReduce implementations to process SPARQL queries over RDF stored in HBase [8,9].

We are only aware of one work on exposing data from column-oriented stores as RDF. Kiran et al. provide a method for generating a SPARQL endpoint, i.e. a standardized RDF query interface, on top of HBase [10]. They map tables to classes, rows to resources, and columns to properties. Their approach do not scale well with increasing numbers of HBase entries, as the results show that the time taken to map HBase data to RDF is in hours for a few million rows [10].

CSV is widely used for publishing tabular data [11]. The CSV on the Web W3C Working Group¹⁴ provides technologies for data dependent applications on the Web working with CSV. Several existing implementations, including that of Anything To Triples (any23)¹⁵, map CSV to a generic RDF representation. Customizable mappings are more suitable for our purpose. In Tarql (Transformation SPARQL)¹⁶, one can define such mappings in SPARQL; Tabels (Tabular

¹⁴ <http://www.w3.org/2013/05/lcsv-charter.html>

¹⁵ <http://any23.apache.org>

¹⁶ <https://tarql.github.io>

Cells)¹⁷ and Sparqlify¹⁸ use domain-specific languages similar to SPARQL. *Tables* provides auxiliary machinery to filter and compare data values during the transformation process. Sparqlify is mainly designed to map relational databases to RDF but also features the sparqlify-csv module.

XML is used for various data and document exchange purposes. Like for CSV→RDF, there are generic and domain-specific XML→RDF approaches. Breitling implemented a direct, schema-independent transformation, which retains the XML structure [13]. Turning this generic RDF representation into a domain-specific one requires post-processing on the RDF side, e.g., transformations using SPARQL CONSTRUCT queries. On the other hand, the current version of Breitling’s approach is implemented in XSLT 1.0, which does not support streaming and is therefore not suitable for the very large inputs of the OpenAIRE setting. Klein uses RDF Schema to map XML elements and attributes to RDF classes and properties [14]. It does not automatically interpret the parent-child relation between two XML elements as a property between two resources, but a lot of such relationships exist in the OpenAIRE XML. XSPARQL can transform XML to RDF and back by combining the XQuery and SPARQL query languages to [15]; authoring mappings requires good knowledge of both. By supporting XQuery’s expressive mapping constructs, XSPARQL requires access to the whole XML input via its DOM (Document Object Model), which results in heavy memory consumption. A subset of XQuery¹⁹ is suitable for streaming but neither supported by the XSPARQL implementation nor by the free version of the Saxon XQuery processor required to run XSPARQL.

6 Implementation

As the only existing **HBase→RDF** implementation does not scale well (cf. section 5), we decided to follow the MapReduce paradigm for processing massive amounts of data in parallel over multiple nodes. We implemented a single MapReduce job. Its mapper reads the attributes and values of the OpenAIRE entities from their protocol buffer serialization and thus obtains all information required for the mapping to RDF. Hence no reducer is required. The map-only approach performs well thanks to avoiding the computationally intensive shuffling. RDF subjects are generated from row keys, predicates and objects from attribute names and cell values or, for linked entities, from column families/qualifiers.

Mapping the OpenAIRE **CSV→RDF** is straightforward: files correspond to classes, columns to properties, and each row is mapped to a resource. We initially implemented mappings in Tarql, Sparqlify and Tables (cf. section 5)

¹⁷ <http://idi.fundacionctic.org/tables>

¹⁸ <https://github.com/AKSW/Sparqlify> [12]

¹⁹ cf. ‘Streaming in XQuery’, <http://www.saxonica.com/html/documentation/sourcedocs/streaming/streamed-query.html>

and ended up preferring Tarql because of its good performance²⁰ and the most flexible mapping language – standard SPARQL²¹ with a few extensions. As we *map* CSV→RDF, as opposed to *querying* CSV like RDF, we implemented *CONSTRUCT* queries, which specify an RDF template in which, for each row of the CSV, variables are instantiated with the cell values of given columns.

To enable easy maintenance of **XML→RDF** mappings by domain experts, and efficient mapping of large XML inputs, we implemented our own approach²². It employs a SAX parser and thus supports streaming. Our mapping language is based on RDF triple templates and on the XPath²³ language for addressing content in XML. XPath expressions in the subjects or objects of RDF triple templates indicate where in the XML they obtain their values from. To keep XPath expressions simple and intuitive, we allow them to be ambiguous, e.g., by saying that *oaf:result/publisher/text()* (referring to the text content of the *publisher* element of a result) maps to the *dterms:publisher* property of an *oav:Result*, and that *oaf:result/dateofacceptance/text()* maps to *dterms:dateAccepted*. In theory, any combination of *publisher* and *dateofacceptance* elements would match such a pattern; however in reality only those nodes that have the shortest distance in the XML document tree represent attributes of the *same* OpenAIRE entity. XML Filters [16] efficiently restrict the XPath expressions to such combinations.

7 Evaluation

7.1 Comparison Metrics

The **time** it takes to transform the complete OpenAIRE input data to RDF is the most important performance metric (requirement R4). The **main memory usage** of the transformation process is important because OpenAIRE2020 envisages the development of further services sharing the same infrastructure, including deduplication, data mining to measure research impact, classification of publications by machine learning, etc. One objective metric for **maintainability** is the size of the mapping’s source code – after stripping comments and compression, which makes the comparison ‘independent of arbitrary factors like lengths of identifiers and amount of whitespace’ [17].²⁴ The ‘cognitive dimensions of notation’ (CD) evaluation framework provides further criteria for systematically assessing the ‘usability of information artefacts’ [18]. The following dimensions are straightforward to observe here: *closeness* of the notation to the problem (here: mapping HBase/CSV/XML to RDF), *terseness* (here measured by code

²⁰ Tabela failed to handle large CSV files because it loads all the data from the CSV into main memory; Sparqlify works similar to Tarql but with almost doubled execution time (7,659 s) and more than doubled memory usage.

²¹ <http://www.w3.org/TR/sparql11-query/>

²² See source code and documentation at <https://github.com/allen501pc/XML2RDF>.

²³ <http://www.w3.org/TR/xpath20/>

²⁴ We used `tar cf - <input files> | xz -9`. For HBase, we considered the part of the Java source code that is concerned with declaring the mapping, whereas our CSV and XML mappings are natively defined in high-level mapping languages.

size; see above), *error-proneness*, *progressive evaluation* (i.e. whether one can start with an incomplete mapping rule and evolve it to further completeness), and *secondary notation and escape from formalism* (e.g. whether reading cues can be given by non-syntactic means such as indentation or comments).

7.2 Evaluation Setup

The **HBase**→**RDF** evaluation ran on a Hadoop cluster of 12 worker nodes operated by CNR.²⁵ As our **CSV**→**RDF** and **XML**→**RDF** implementations required dependencies not yet installed there, we evaluated them locally: on a virtual machine on a server with an Intel Xeon E5-2690 CPU, having 3.7 GB memory and 250 GB disk space assigned and running Linux 3.11 and JDK 1.7. As we did not have a cluster available, and as the tools employed did not natively support parallelization, we ran the mappings from CSV and XML sequentially.

7.3 Measurements and Observations

The following table lists our measurements; further observations follow below.

Objective Comparison Metrics	HBase	CSV	XML
Mapping Time(s)	1,043	4,895	45,362
Memory (MB)	68,000	103	130
Compressed Mapping Source Code (KB)	4.9	2.86	1.67
Number of Input rows/records	20,985,097	203,615,518	25,182,730
Number of Generated RDF Triples	655,328,355	654,193,273	788,953,122

For **HBase**→**RDF**, the peak memory usage of the cluster was 68 GB, i.e. ~ 5.5 GB per worker node. No other MapReduce job was running on the cluster at the same time; however, the usage figure includes the memory used by the Hadoop framework, which schedules and monitors job execution.

The 20 **CSV** input files correspond to different entities but also to relationships. This, plus the way multi-valued attributes are represented (cf. section 2.2), causes the high number of input rows. The size of all files is 33.8 GB. The **XML**→**RDF** memory consumption is low because of stream processing. The time complexity of our mapping approach depends on the number of rules (here: 118) and the size of the input (here: 144 GB). With the complexity of the XML representation, this results in an execution time of more than 12 hours. The size of the single RDF output file is ~ 91 GB. Regarding *cognitive dimensions*, the different notations expose the following characteristics; for lack of space we focus on selected highlights. *Terseness*: the high-level CSV→RDF and XML→RDF languages fare better than the Java code required for HBase→RDF. Also, w.r.t. *closeness*, they enable more intuitive descriptions of mappings. As the CSV→RDF mappings are based on SPARQL, which uses the same syntax for RDF triples than the Turtle RDF serialization, they look

²⁵ https://issue.openaire.research-infrastructures.eu/projects/openaire/wiki/Hadoop_Clusters#section-3

closest to RDF. *Error-proneness*: Syntactically correct HBase→RDF Java code may still define a semantically wrong mapping. In Tarql’s CSV→RDF mappings, many types of syntax and semantics errors can be detected easily. *Progressive evaluation*: one can start with an incomplete Tarql mapping rule CSV→RDF mapping rule and evolve it towards completeness. *Secondary notation*: Tarql and Java support flexible line breaks, indentation and comments, whereas our current XML→RDF mapping implementation requires one (possibly long) line per mapping rule. Overall, this strongly suggests that CSV→RDF is the most maintainable approach.

8 Related Work

Comparisons of different approaches of mapping data to RDF have mainly been carried out for relational databases as a source [19,?]. Similarly to our evaluation criteria, the reference comparison framework of the W3C RDB2RDF Incubator Group covers mapping creation, representation and accessibility, and support for data integration [21]. Hert et al. compared different RDB2RDF mapping languages w.r.t. syntactic features and semantic expressiveness [22].

For other linked datasets about research, we refer to the ‘publication’ and ‘government’ sectors of the LOD Cloud, which comprises, e.g., publication databases such as DBLP, as well as snapshots of funding databases such as CORDIS. From this it can be seen that OpenAIRE is a more comprehensive data source than those published as LOD before.

9 Conclusion and future work

We have mapped a recent snapshot of the OpenAIRE data to RDF. A preliminary dump as well as the definitions of the mappings are available online at <http://tinyurl.com/OALOD>. Mapping from HBase is fastest, whereas mapping from CSV promises to be most maintainable. Its slower execution time is partly due to the less powerful hardware on which we ran it; comparing multiple CSV→RDF processes running in parallel to the HBase→RDF implementation on the CNR Hadoop cluster seems promising. Based on these findings the OpenAIRE2020 LOD team will decide on the preferred approach for providing the OpenAIRE data as LOD; we will then make the data available for browsing from their OpenAIRE entity URIs, and for querying via a SPARQL endpoint.

Having implemented almost the whole OpenAIRE data model, future steps include interlinking the output with other existing datasets. E.g., we so far output countries and languages as strings, whereas DBpedia and Lexvo.org are suitable linked open datasets for such terms. Link discovery tools will further enable large-scale linking against existing ‘publication’ and ‘government’ datasets.

Acknowledgements. We would like to thank the partners in the OpenAIRE2020 project, in particular Claudio Atzori, Alessia Bardi, Glykeria Katsari and Paolo Manghi, for their help with accessing the OpenAIRE data. This work has been partially funded by the European Commission under grant agreement no. 643410.

References

1. F. Osborne and E. Motta, "Understanding research dynamics," in *Semantic Web Evaluation Challenges 2014*, no. 457 in CCIS, Springer, 2014.
2. F. Scharffe, G. Atemezing, R. Troncy, F. Gandon, S. Villata, B. Bucher, F. Hamdi, L. Bihanic, G. Képékian, and F. Cotton, "Enabling linked data publication with the Datalift platform," in *Proc. AAAI workshop on semantic cities*, 2012.
3. P. Manghi, M. Mikulicic, and C. Atzori, "Openaire data model specification," deliverable.
4. P. Manghi, N. Houssos, M. Mikulicic, and B. Jörg, "The data model of the openaire scientific communication e-infrastructure," in *Metadata and Semantics Research*, Springer, 2012.
5. "Protocol buffers," 2015.
6. J. Sun and Q. Jin, "Scalable rdf store based on hbase and mapreduce," in *Advanced Computer Theory and Engineering (ICACTE)*, vol. 1, IEEE, 2010.
7. V. Khadilkar, M. Kantarcioglu, B. Thuraisingham, and P. Castagna, "Jena-hbase: A distributed, scalable and efficient rdf triple store," in *ISWC Posters & Demonstrations*, 2012.
8. N. Papailiou, I. Konstantinou, D. Tsoumakos, and N. Koziris, "H2rdf: adaptive query processing on rdf data in the cloud.," in *International conference on World Wide Web*, ACM, 2012.
9. A. Haque and L. Perkins, "Distributed rdf triple store using hbase and hive," *University of Texas at Austin*, 2012.
10. K. V. K. and D. G. S. Sadasivam, "A novel method for dynamic sparql endpoint generation in nosql databases," *Australian Journal of Basic and Applied Sciences*, vol. 9, no. 6, 2015.
11. T. Lebo and G. T. Williams, "Converting governmental datasets into linked data," in *I-Semantics*, ACM, 2010.
12. I. Ermilov, S. Auer, and C. Stadler, "Csv2rdf: User-driven csv to rdf mass conversion framework," in *I-Semantics*, 2013.
13. F. Breitling, "A standard transformation from xml to rdf via xslt," *Astronomische Nachrichten*, vol. 330, no. 7, 2009.
14. M. Klein, "Interpreting xml documents via an rdf schema ontology," in *DEXA*, 2002.
15. S. Bischof, S. Decker, T. Krennwallner, N. Lopes, and A. Polleres, "Mapping between rdf and xml with xsparql," *Journal on Data Semantics*, vol. 1, no. 3, 2012.
16. Y. Diao, P. Fischer, M. J. Franklin, and R. To, "Yfilter: Efficient and scalable filtering of xml documents," in *Data Engineering*, IEEE, 2002.
17. F. Wiedijk, "The "de Bruijn factor"," 2012.
18. A. Blackwell and T. Green, "Cognitive dimensions of notations resource site," 2010.
19. M. Svihla and I. Jelinek, "Benchmarking rdf production tools," in *DEXA*, Springer, 2007.
20. F. Michel, J. Montagnat, and C. Faron-Zucker, "A survey of rdb to rdf translation approaches and tools," 2014.
21. S. S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. T. Jr, S. Auer, J. Sequeda, and A. Ezzat, "A survey of current approaches for mapping of relational databases to rdf," 2009.
22. M. Hert, G. Reif, and H. C. Gall, "A comparison of rdb-to-rdf mapping languages," in *I-Semantics*, ACM, 2011.