ALIGNED

Aligned, Quality-centric, Software and Data Engineering

H2020 – 644055

D7.12 - Data Management Plan

Project Number:	644055
Project Title:	ALIGNED
Document Type: (Deliverable/Internal):	Deliverable
Deliverable Number:	D7.12
Deliverable Type (R/DEM/DEC/OTHER):	REPORT
Workpackages Contributing:	WP7
Dissemination Level (PU/CO):	PU - Public
Contractual Delivery Date:	January 2017
Actual Delivery Date:	January 2017
Version:	1.10
Editor(s)/Lead Authors:	Dimitris Kontokostas, Markus Freudenbeurg
With contributions from:	Christian Dirschl, Bojan Božić, Rob Brennan, Monika Solanki, Gavin Mendel-Gleason, Rob Brennan, Martin Brümmer, Andreas Koller
Reviewers(s):	Monika Solanki, Kevin Feeney

Abstract:

This report provides the final version of the Data Management Plan for the ALIGNED project. It describes how the project data will be published as Linked Data as part of the H2020 open research data pilot.

Keyword List:

Data Management, DataID, Open Data, RDF, Linked Data, Open Data Pilot

History

Version	Date	Reason	Revised by
1.01	1/5/2015	First Draft (D7.2)	Dimitris Kontokostas
1.02	11/5/2015	Reviewer comments	Monika Solanki
1.03	25/5/2015	Revised Internal Draft for Final Review	Dimitris Kontokostas
1.04	25/5/2015	Final Version (D7.2)	Rob Brennan
1.05	15/11/2015	Review comments (D7.11 input)	Rob Brennan
1.06	06/01/2016	First Draft (D7.11)	Dimitris Kontokostas
1.07	25/01/2016	Final version (D7.11)	Rob Brennan
1.08	06/01/2017	First Draft (D7.12)	Dimitris Kontokostas
1.09	20/01/2017	Reviewer comments	Monika Solanki
1.10	30/01/2017	Final version (D7.12)	Kevin Feeney

Executive Summary

This deliverable describes all of the data generated, collected and curated within the ALIGNED project. It describes how we use DataID as a machine readable metadata format to describe ALIGNED datasets as well as existing ALIGNED meta model vocabularies. This is the final iteration and reports only the updates from the previous version.

Table of Contents

1 Introduction

This deliverable describes all of the data generated, collected and curated within the ALIGNED project. For the dataset descriptions we used the official guidelines in the Data Management Plan (DMP) from the Horizon 2020 portal[1]. The current deliverable the final iteration of the data management plan but we already moved all the dataset description on the website, which, from now on will be the main reference during the lifespan of the ALIGNED project. Changes in this deliverable since D7.11 are provided in section 1.1.

As previously reported, the contents of this deliverable will inform the work of all ALIGNED work packages but especially WP2 (Meta-model Requirements and Design) where the publication of the ALIGNED meta-model as Linked Data vocabularies will follow the patterns laid out here and WP6 (Validation) where experimental data will be collected and published using these guidelines. We continue to use DataID as the basis for dataset meta-data and the use of the *DataID to DMP documentation generator*) tool. A tool that takes an DataID based meta-data description in RDF and renders it into a human-readable format.

The structure of the rest of this document is as follows: Chapter 2 describes the ALIGNED Data Management Plan approach, DataID and the DataID to DMP tool. Chapter 3 provides dataset and metamodel descriptions and Chapter 4 reviews future work.

1.1 Changes since deliverable D7.11

The following updates exist in this deliverable in comparison to D7.11
- Minor updates to section 2
- Overview of updated DataID model
- Updated data descriptions for DBpedia, Seshat and JURION
- New dataset description for PoolParty

2 ALIGNED Data Management Plan

The ALIGNED project is focused on RDF and Linked Data. RDF is a W3C standard and is already surrounded by a complete software stack for reading, storing, querying, analysing, assessing, curating and enriching RDF data. Thus, the RDF data provided by ALIGNED provide out-of-the box compliance to the additional guidance (Annex 2) of the DMP guidelines (discoverability, accessibility, assessability and interoperability). Regarding usability beyond the original purpose, it is described on a per-dataset basis. Ethics procedures govern all personal data.

For ALIGNED datasets we use DataID [1] as a machine readable metadata format based on RDF for all the data collected, processed or generated within ALIGNED. A detailed description of

---

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (accessed May 15th 2015)

DataID with ADMS[2] (Asset Description Metadata Schema) was provided in the previous version of this deliverable and we see DataID as complementary to ADMS.

ALIGNED ontologies or vocabularies, e.g. ALIGNED meta models or partner ontologies, use Persistent URLs (PURLs) that with Linked Data provide RDF for machines or HTML for humans through content negotiation. The project website aggregates all defined ontologies or vocabularies.

2.1 DataID

Since the inception of DataID [1], this ontology has developed into a more generic and far reaching approach to portray dataset metadata. Due to a plethora of use case specific requirements, the monolithic ontology of a first version was modularised into multiple extension ontologies around a single core, establishing the DataID Ecosystem (Figure 1). We demonstrated the capabilities of the DataID Ecosystem [2] to accommodate for complex metadata requirements (i.e. Data Management Plans), expanding DataID core with existing extensions, while maintaining the ability to interoperate with other metadata vocabularies in turn. Furthermore, the DataID Ecosystem shall provide a controlled environment, in which the interplay of different extensions and DataID core will be regulated and documented, making it easy to combine multiple extensions with DataID core to solve a given usage scenario.
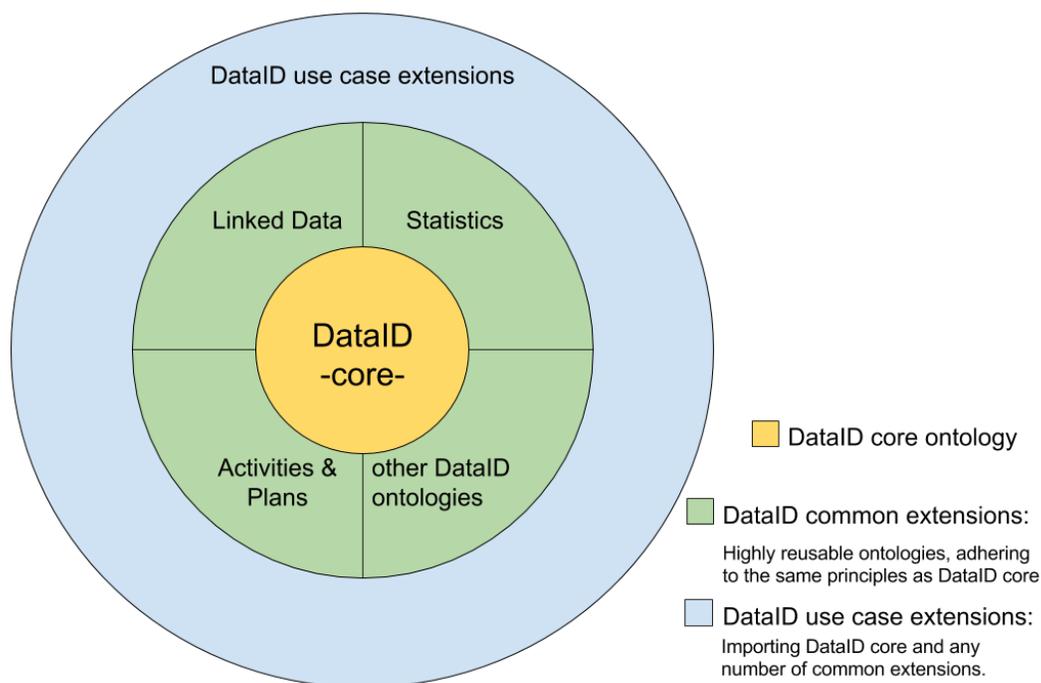

Figure 1: The DataID Ecosystem

---

[2] http://www.w3.org/TR/vocab-adms/ (accessed May 15th 2015)

Together with the Data Lifecycle Ontology (DLO), DataID core forms the data engineering side of the ALIGNED Suite of Ontologies [3], describing datasets and their different manifestations, in the context of a given domain, as well as relations to agents like persons or organisations, in regard to their rights and responsibilities. The requirements for the DMP of the ALIGNED project are now covered entirely by DataID core without any further extension, since all of the textual statements (on openness, reuse and integration, etc.) introduced by the original extension for DMPs (cf. deliverable D7.11), are relevant information for dissemination tasks, independent of the context of a Data Management Plan.

2.2 DataID - DMP Generator

For the purpose of this deliverable we created a DataID to DMP generator[3]. The generator takes as input a DataID file and produces an HTML report. We upload both the DataIDs and the HTML reports on the ALIGNED website.

3 ALIGNED Data Descriptions

In the following sections we describe data that partners already have and plan to use in ALIGNED.

3.1 ALIGNED Datasets

Compared to the previous deliverable iteration, we do not inline the metadata descriptions of our datasets but instead, refer to the ALIGNED website for both machine & human readable versions.

ALIGNED provides the following datasets:

DBpedia (updated)
- Human readable: http://aligned-project.eu/dataids/DBpedia.html
- Machine readable: http://aligned-project.eu/dataids/DBpedia.ttl

Seshat (updated)
- Human readable: http://aligned-project.eu/dataids/Seshat.html
- Machine readable: http://aligned-project.eu/dataids/Seshat.ttl

JURION (updated)
- Human readable: http://aligned-project.eu/dataids/Jurion.html
- Machine readable: http://aligned-project.eu/dataids/Jurion.ttl

PoolParty (new)
- Human readable: http://aligned-project.eu/dataids/PoolParty-UG.html
- Machine readable: http://aligned-project.eu/dataids/PoolParty-UG.ttl

---

[3] https://github.com/dbpedia/dataid/tree/master/dmp (accessed May 15th 2015)

3.4 ALIGNED Meta Model Vocabularies

All the ALIGNED meta model vocabularies are listed on the project website[4] both with a human readable form (html) as well as a machine readable one (RDF) as Linked Data. As described in Deliverable 6.2 APPENDIX 6, the meta model vocabularies publishing process is described as follows.

The metamodel is developed as a set of self-documenting RDF vocabularies that are suitable for consumption by the LODE tool[5]. The canonical development source for change-controlled ALIGNED metamodel vocabularies is a public Github repository[6]. The canonical publishing location for metamodel vocabularies, documentation, metadata and linked data is on the ALIGNED project website. The development and publishing process used is illustrated as a flow-chart in figure 2. A description of each step is provided in the previous iteration of this deliverable.
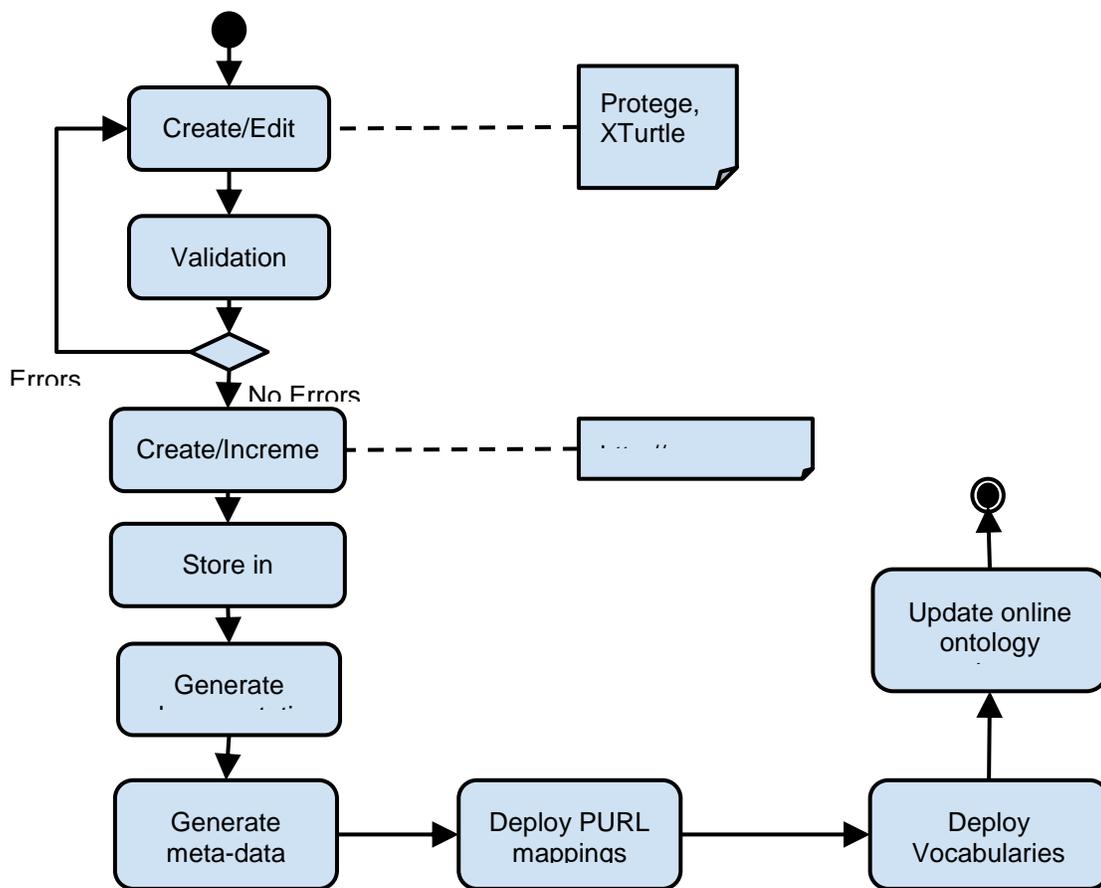

Figure 2: ALIGNED Metamodel Vocabulary Publishing Process

3.5 ALIGNED experimental datasets

Up to this time, no partner keeps any ALIGNED related experimental datasets. As scientific papers are published that reference experimental data, the data will be made accessible through the ALIGNED website.

3.6 ALIGNED-related data & metadata

Besides the ALIGNED datasets and vocabularies, the project keeps other up-to-date project related information such as:

- **Publications**: aggregated at the projects Bibsonomy account[7]

- **Videos**: stored at the project's YouTube account[8]

All the ALIGNED-related information is available from the project website.

4 Future Work

The ALIGNED project website already acts as a centralised hub of meta-data for the project datasets so that a visitor can find all published data. The project website will remain up to date with ALIGNED dataset descriptions for the lifespan of the ALIGNED project.

5 References

[1] M. Brummer, C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, and S. Hellmann. DataID: Towards semantically rich metadata for complex datasets. In Proc. of the 10th International Conference on Semantic Systems, pages 84–91. ACM, 2014.

[2] M. Freudenberg, M. Brümmer, J. Rücknagel, R. Ulrich, T. Eckart, D. Kontokostas and S. Hellmann "The Metadata Ecosystem of DataID". In: Metadata and Semantics Research: 10th International Conference, MTSR 2016, Göttingen, Germany.

[3] M. Solanki, B. Bozic, M. Freudenberg, D. Kontokostas, C. Dirschl and R. Brennan "Enabling Combined Software and Data Engineering at Web-Scale: The ALIGNED Suite of Ontologies". In: The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan.

---

[7] http://www.bibsonomy.org/group/aligned/aligned-project
[8] https://www.youtube.com/channel/UCRJAHer4rhNwlPnZO5Dastg