D2.3 - DATA LINKING TECHNOLOGIES

Co-funded by the Horizon 2020
Framework Programme of the European Union

DELIVERABLE NUMBER	D2.3
DELIVERABLE TITLE	Data Linking Technologies
RESPONSIBLE AUTHOR	Agroknow

GRANT AGREEMENT N.	731001
PROJECT ACRONYM	AGINFRA PLUS
PROJECT FULL NAME	Accelerating user-driven e-infrastructure innovation in Food & Agriculture
STARTING DATE (DUR.)	01/01/2017 (36 months)
ENDING DATE	31/12/2019
PROJECT WEBSITE	plus.aginfra.eu
COORDINATOR	Nikos Manouselis
ADDRESS	17 Grammou Str., Vrilissia GR15235, Greece
REPLY TO	nikosm@agroknow.com
PHONE	+30 210 6897 905
EU PROJECT OFFICER	Mrs. Pilar Ocon-Garces

WORKPACKAGE N. | TITLE	WP2 | Data & Semantics Layer
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D2.3 | Data Linking Technologies
RESPONSIBLE AUTHOR	Antonis Koukourikos
REPLY TO	akukurik@agroknow.com
DOCUMENT URL	http://www.plus.aginfra.eu/sites/plus_deliverables/D2.3.pdf
DATE OF DELIVERY (CONTRACTUAL)	30/09/2017 (M9)
DATE OF DELIVERY (SUBMITTED)	29/09/2017 (M9)
VERSION | STATUS	v1.0 | First submission to the EC
NATURE	P (Prototype)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Pythagoras Karampiperis, Antonis Koukourikos (Agroknow)
REVIEWERS	Teodor Georgiev (PENSOFT)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Preliminary Tools and Methods review	31/05/2017	Agroknow
0.3	Harmonization with Requirements	31/07/2017	Agroknow
0.5	Silk framework assessment	07/09/2017	Agroknow
0.6	Report setup	15/09/2017	Agroknow
0.7	Report draft finalization	22/09/2017	Agroknow
0.8	Deliverable Review	27/09/2017	PENSOFT
0.9	Deliverable finalization	28/09/2017	Agroknow
1.0	Submission to the EC	29/09/2017	Agroknow

PARTICIPANTS		CONTACT
Agro-Know IKE (Agroknow, Greece)		Nikos Manouselis Email: nikosm@agroknow.com
Stichting Wageningen Research (DLO, The Netherlands)		Rob Lokers Email: rob.lokers@wur.nl
Institut National de la Recherché Agronomique (INRA, France)		Pascal Neveu Email: pascal.neveu@inra.fr
Bundesinstitut für Risikobewertung (BFR, Germany)		Matthias Filter Email: matthias.filter@bfr.bund.de
Consiglio Nazionale Delle Richerche (CNR, Italy)		Leonardo Candela Email: leonardo.candela@isti.cnr.it
University of Athens (UoA, Greece)		George Kakaletris Email: gkakas@di.uoa.gr
Stichting EGI (EGI.eu, The Netherlands)		Tiziana Ferrari Email: tiziana.ferrari@egi.eu
Pensoft Publishers Ltd (PENSOFT, Bulgaria)		Lyubomir Penev Email: penev@pensoft.net

ACRONYMS LIST

API	Application Programming Interface
JVM	Java Virtual Machine
LOD	Linked Open Data
LSL	Link Specification Language
OWL	Web Ontology Language
RDF	Resource Description Framework
REST	Representational state transfer
SKOS	Simple Knowledge Organisation System
SPARQL	SPARQL Protocol and RDF Query Language
VRE	Virtual Research Environment

EXECUTIVE SUMMARY

The present report is the first submitted iteration of a living document that will describe progress and evolvement of the AGINFRA PLUS data linking components, i.e. the services that will be incorporated in the overall AGINFRA PLUS architecture and be responsible for providing indicative links and relations between heterogeneous data assets, e.g. inclusion and relevance relations between publications, datasets and models, similarities in conceptualizations, etc.

The current version of the deliverable focuses on the description of the core linking techniques commonly used for semantically rich data assets, as well as, the installation and deployment of a major data linking platform to serve as a showcase and baseline of the functionalities to be introduced in AGINFRA PLUS. It is expected that the usage of the data linking technologies will be significantly customized to the needs of the involved research communities, and will be tailored to their needs as well as their usability and ease-of-use requirements from a relatively complex tool.

It is expected that, as the use cases are refined and executed, the data linking components will be accordingly updated and extended or modified. Additionally, general developments on the used baseline tools will be monitored and adopted if suitable for the purposes of AGINFRA PLUS. To this end, the report is treated as a living document, with regular submission to the EC of versions that report on significant changes in the respective prototypes.

TABLE OF CONTENTS

TABLE OF FIGURES

1 INTRODUCTION

In general terms, data linking is the task of determining whether two object descriptions can be linked one to the other with a type of relation that holds between them in order to represent the fact that they refer to the same real-world object in a specific domain. Quite often, this task is performed on the basis of the evaluation of the degree of similarity among different data instances describing real-world objects across heterogeneous data sources, under the assumption that the higher is the similarity between two data descriptions, the higher is the probability that the two descriptions actually refer to the same object. From an operational point of view, data linking also refers to the tasks and processes of defining methods, techniques and (semi-)automated tools for performing the similarity evaluation task.

Data linking can be formalized as an operation which takes two collections of data as input and produces a collection of mappings between entities of the two collections as output. Mappings denote binary relations between entities corresponding semantically one to another.

In the context of the Semantic Web, data linking is materialized via the Linked Data Initiative[1], which calls for datasets to provide links to other published resources, thus building the continuously expanding Linked Data Cloud[2].

While the majority of currently established data links pertain to identity or hierarchical relationships, in theory a data linking process can define arbitrary relations between the datasets to be linked. Furthermore, as notions like identity, inclusion and similarity are to some extent subjective and the subject of largely philosophical and interpretational debate, a practical system has to rely on certain assumptions for reaching a decision on a relation between two entities.

Summarizing the aforementioned points, the following figure showcases the general workflow for a semi-automatic data linking service.

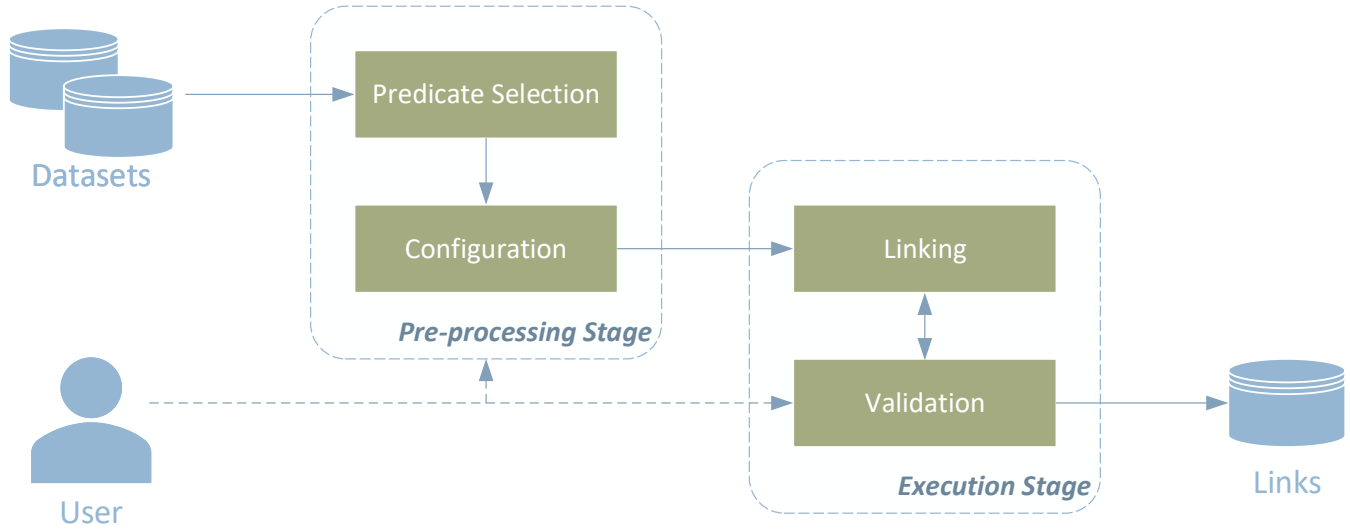

Figure 1: Generic Data Linking Workflow

The pre-processing stage of the workflow includes a *Predicate Selection* process, where the relations that will be used for linking the examined datasets are specified.

It also incorporates a *Configuration* stage, where the parameters for executing the actual linking process are set. The range and means of configuration can vary widely between linking systems depending on their adopted approach for finding links. For example, it may be required to determine thresholds for the different supported predicates in order for the system to decide if a relation holds. Similarly, depending on the nature of the compared data, the user may have to decide on the specific comparison methodologies to be employed for the link discovery process.

[1] http://linkeddata.org/

[2] http://lod-cloud.net/

At the *Linking* stage, the methods that carry out the link discovery process are executed in accordance with the parameterizations defined in the pre-processing stage. As mentioned, the Linking component may employ more than one matching technique, and may use external resources for actually deciding on the validity of a relation between two entities in the processed datasets.

The links proposed by the automatic phase of the execution stage are usually subject to examination and *Validation* from a human expert. The provision of human judgement can actually be used to further inform the automatic system for subsequent (or independent future) runs, so the Linking and Validation stage may constitute a loop in the workflow, terminated at the decision of the user.

In the following subsections of the report, we identify the linking requirements of the AGINFRA PLUS framework, discuss on the specification of a linking system that will be able to serve them, and devise an action plan for carrying out the required steps towards establishing the platform's data linking components, as well as, the workflow of their usage within the designed use cases.

2 AGINFRA PLUS DATA LINKING REQUIREMENTS

To understand the data linking requirements of AGINFRA PLUS, we must consider the range and nature of data assets that will be handled within the framework. Based on the requirements analysis reported in deliverable D2.1, we focus on three types of data assets:

- Models;
- Datasets;
- Publications.

As these three types are clearly linked within the overall flow of scientific research design, execution and dissemination, it follows that we need two core categories of linking relevant to AGINFRA PLUS:

- Relations between data assets of the same type;
- Relations between data assets of different types.

That is, we need to identify or establish predicates for defining relations between entities of the same type, as well as predicates for expressing relations between entities of different data types (e.g. that a publication refers to a specific model or, more intricately, that the work described in a publication uses a model comparable to a different model found in other publications or repositories).

This essentially means that we cannot be restricted to systems and methods that are limited to using the dominant *owl:sameAs* predicate (which indicates solely identity) or even the "match" set of predicates defined in SKOS (e.g. *broaderMatch*, *narrowerMatch*, *closeMatch*), but we need the ability to define different predicates for linking different entities.

Another important point to consider is the types of linking approaches that are applicable to the data assets handled within AGINFRA PLUS. While basic string and value matching techniques applied over the metadata descriptions of data assets produce generally good results with good efficiency, the need for the discovery of more complex relations likely leads to the incorporation of knowledge-based comparison methodologies, which take into account structural characteristics of the metadata schema used for describing the examined data assets and, furthermore, activate external knowledge sources.

An additional parameter is that the general classification of value matching techniques is not adequate for determining which method is actually most suitable for a given dataset comparison task. Depending on the expected variations on expressivity and method of metadata annotation (manual, extracted, generated), the use of a Levenshtein edit distance metric ([1], [2], [3]) can be more accurate than a Jaro metric ([4], [5], [6]) or vice-versa. Furthermore, string or linguistic comparisons may prove insufficient for different dataset comparisons, thus other techniques using more complex approaches, like rule-based structure analysis [7], mapping of axiomatic relations found in the data schema [8] in valuation networks [9], belief propagation approaches [10], etc., are adopted. It is, therefore, necessary to design a linking system that allows the inclusion of different methodologies that will be activated on occasion, depending on the nature of a specific linking task.

3 PROPOSED BASELINE FRAMEWORK

While there are multiple data linking frameworks proposed in the literature ([11], [12], [13], [14]), they are mostly research prototypes with inadequate stability for incorporation in an operational framework, and consequent lack of support and maintenance.

The additional requirements on flexibility and relative ease of configuration and execution constitute the Silk framework[3] [1] a good candidate solution to serve as the baseline platform for building the AGINFRA PLUS data linking components.

As most frameworks, Silk supports the generation of *owl:sameAs* links but allows the declaration of arbitrary RDF predicates to be used as the link relation between connected entities. Furthermore, its implementation in Scala and therefore the ability to be executed over a standard JVM suits itself to incorporation in the overall VRE infrastructure adopted by AGINFRA PLUS. The framework also provides a REST API for incorporating its functionalities within the overall workflows that will be designed to run over the VRE.

To determine link relations and link conditions, Silk uses its own XML-based formalism, the *Link Specification Language* (Silk-LSL). A Silk-LSL record defines access parameters for the involved dataset descriptions (that must be accessible via SPARQL endpoints), execution parameters for the Silk framework, link identification metrics to be activated and aggregation functions for combining these metrics. Regarding the metrics used by Silk, they are predominantly focused on string matching methodologies, however, the framework allows the incorporation of external modules implementing different metrics for comparing the entities within the compared datasets.

The demo installation that will be used for AGINFRA PLUS is deployed at: http://83.212.101.14:9000

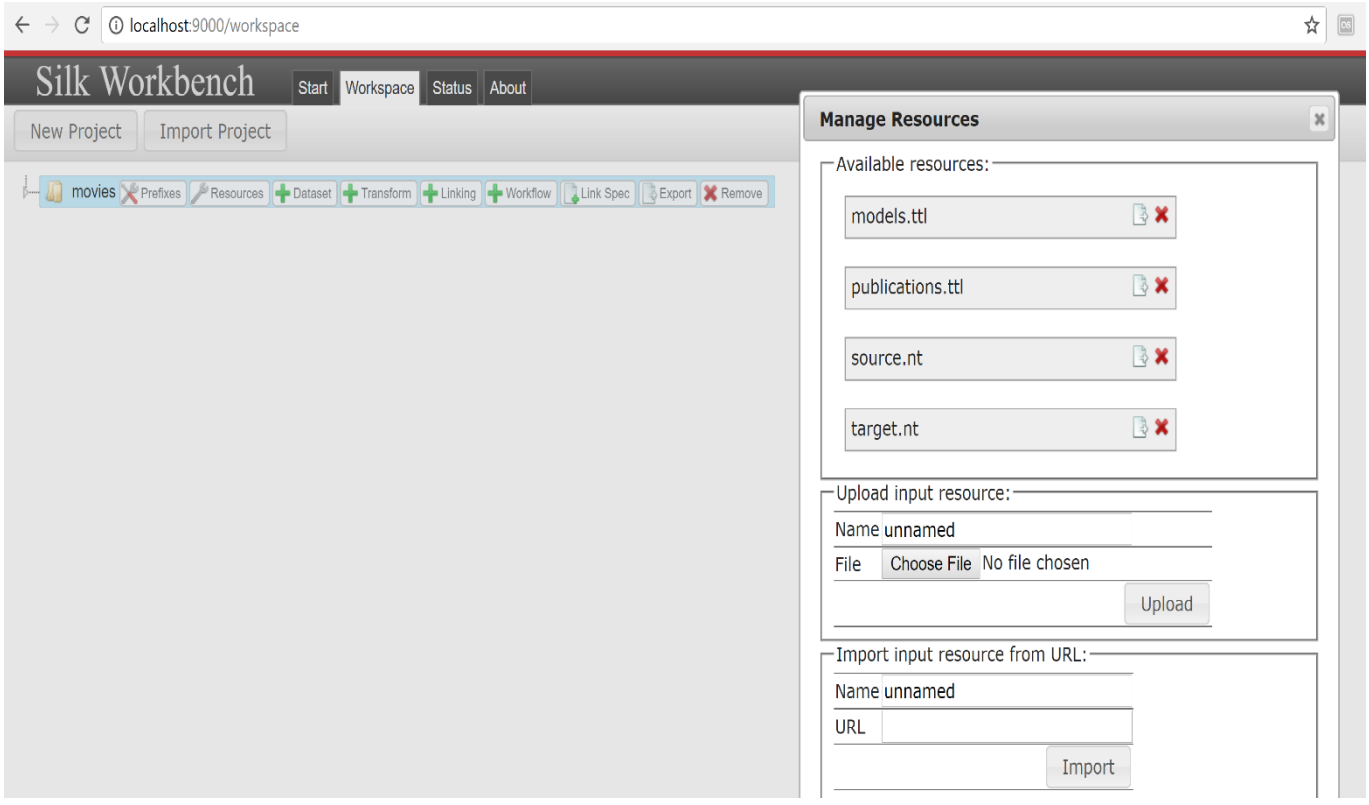

Figure 2: Input Dataset Management in Silk

[3] http://silkframework.org/

```
<Silk>
 <DataSource id="publications">
  <EndpointURI>http://83.212.101.14:7200/aginfra-publications/sparql</EndpointURI>
  <Graph>http://83.212.101.14:7200/aginfra-publications</Graph>
  <DoCache>1</DoCache>
  <PageSize>1000</PageSize>
 </DataSource>
 <DataSource id="models">
  <EndpointURI>http://83.212.101.14:7200/aginfra-models/sparql</EndpointURI>
  <Graph>http://83.212.101.14:7200/aginfra-models</Graph>
  <DoCache>1</DoCache>
  <PageSize>1000</PageSize>
 </DataSource>
 <Metric id="jaroSets">
  <Param name="item1" />
  <Param name="item2" />
  <AVG>
   <Compare metric="jaroWinklerSimilarity">
    <Param name="str1" path="?item1" />
    <Param name="str2" path="?item2" />
   </Compare>
  </AVG>
 </Metric>
 <Interlink id="usage">
  <LinkType>aginfra:appliesTo</LinkType>
  <SourceDataset dataSource="publications" var="a" />
  <TargetDataset dataSource="models" var="b" />
  <LinkCondition>
   <MAX weight="1">
    <Compare metric="maxSimilarityinSets">
     <Param name="set1" path="?a/rdfs:label" />
     <Param name="set2" path="?b/rdfs:label" />
     <Param name="submetric" value="jaroSets" />
    </Compare>
   </MAX>
  </LinkCondition>
  <Thresholds accept="0.8" verify="0.6" />
  <Output acceptedLinks="output_links.ttl" verifyLinks="check_links.ttl" format="turtle" mode="truncate" />
 </Interlink>
</Silk>
```

Figure 3: Exemplary Silk-LSL File Structure

4 NEXT STEPS AND ACTION PLAN

The first step towards establishing a powerful and efficient AGINFRA PLUS data linking service is the production of a formal specification for the inter-datatype and intra-datatype relations that are critical to support. The action culminates to the production of an upper-level ontology for describing the AGINFRA PLUS data assets and their connections.

Following the establishment of the upper-level ontology, a testing phase for the data linking services will be initiated. During this stage, an initial set of datasets to be linked will be selected and the corresponding Silk-LSL configuration will be constructed. The system's result will be examined by technical and community experts to determine the efficiency of the system and decide on the further improvements to be made (introduction of additional link evaluation methods, changes on data types relations, etc.).

Additionally, work on identifying the requirements for introducing the extended / modified Silk framework in the AGINFRA PLUS VRE will be carried out by the partners involved in the relevant Work Packages.

5 REFERENCES

[1] J. Volz, C. Bizer, M. Gaedke and G. Kobilarov, "Discovering and maintaining links on the web of data," in *8th International Semantic Web Conference (ISWC 2009)*, Washington DC, USA, 2009, http://dx.doi.org/10.1007/978-3-642-04930-9_41.

[2] V. Lopez, A. Nikolov, M. Fernandez, M. Sabou, V. Uren and E. Motta, "Merging and ranking answers in the Semantic Web," in *4th Asian Semantic Web Conference (ASWC)*, Shanghai, China, 2009, http://dx.doi.org/10.1007/978-3-642-10871-6_10.

[3] J. Li, J. Tang, Y. Li and Q. Luo, "RiMOM: A dynamic multistrategy ontology alignment framework," *IEEE Transactions on Knowledge and Data Engineering,* vol. 21, no. 8, pp. 1218-1232, 2009, http://dx.doi.org/10.1109/tkde.2008.202.

[4] O. Udrea, L. Getoor and R. J. Miller, "Leveraging data and structure in ontology integration," in *2007 ACM SIGMOD international conference on Management of Data*, Beijing, China, 2007, http://dx.doi.org/10.1145/1247480.1247531.

[5] E. Ioannou, C. Niederée and W. Nejdl, "Probabilistic Entity Linkage for heterogeneous information spaces," in *20th International Conference on Advanced Information Systems Engineering (CAiSE 2008)*, Montpellier, France, 2008, http://dx.doi.org/10.1007/978-3-540-69534-9_41.

[6] A. Nikolov, V. Uren, E. Motta and A. d. Roeck, "Integration of semantically annotated data by the KnoFuss architecture," in *16th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2008)*, Acitrezza, Italy, 2008, http://dx.doi.org/10.1007/978-3-540-87696-0_24.

[7] F. Säis, N. Pernelle and M.-C. Rousset, "Combining a logical and a numerical method for data reconciliation," *Journal of Data Semantics,* vol. 28, pp. 66-94, 2008, http://dx.doi.org/10.1007/978-3-642-00685-2_3.

[8] A. Nikolov, M. d'Aquin and E. Motta, "Unsupervised instance coreference resolution using a genetic algorithm," 2011.

[9] P. Shenay, "Valuation-based systems: a framework for managing uncertainty in expert systems," in *Fuzzy logic for the management of uncertainty*, New York, NY, USA, John Wiley & Sons, 1992, pp. 83-104.

[10] J. Noessner, M. Niepert, C. Meilicke and H. Stuckenschmidt, "Leveraging terminological structure for object reconciliation," in *7th Extended Semantic Web Conference*, Heraklion, Crete, Greece, 2010, http://dx.doi.org/10.1007/978-3-642-13489-0_23.

[11] W. Hu, J. Chen and Y. Qu, "A self-training approach for resolving object coreference on the semantic web," in *20th International World Wide Web Conference (WWW 2011)*, Hyderabad, India, 2011, http://dx.doi.org/10.1145/1963405.1963421.

[12] Y. Raimond, C. Sutton and M. Sandler, "Automatic interlinking of music datasets on the Semantic Web," in *Linking Data on the Web Workshop at WWW 2008 (LDOW 08)*, Beijing, China, 2008, http://dx.doi.org/10.1109/mmul.2009.29.

[13] A.-C. N. Ngomo and S. Auer, "LIMES - a time-efficient approach for large-scale link discovery on the web of data," in *International Joint Conference on Artificial Intelligence*, Barcelona, Spain, 2011.

[14] X. Niu, S. Rong, Y. Zhang and H. Wang, "Zhishi.links results for OAEI 2011," in *Ontology Matching Workshop at ISWC 2011*, Bonn, Germany, 2011.