EU Project No:601043(Integrated Project (IP))

DIACHRON

Managing the Evolution and Preservation of the Data Web DIACHRON

Dissemination level:	Public
Type of Document:	Documentation
Contractual date of delivery:	36
Actual Date of Delivery:	36
Deliverable Number:	10.7
Deliverable Name:	Standardisation Activities
Deliverable Leader:	UBONN
Work package(s):	WP10
Status & version:	1.0
Number of pages	
WP contributing to the deliverable	WP10
WP / Task responsible	Task 10.5
Coordinator (name / contact)	Jeremy Debattista (UBONN)
Other Contributors	Christoph Lange (UBONN), Giorgos Flouris (FORTH), Hasapis Panagiotis (Intrasoft)
EC Project Officer	Federico Milani
Keywords:	Standardisation Activities
Abstract (few lines):	This document describes the standardisation activities undergone in the DIACHRON Project. We describe how various aspects of the DIACHRON framework adopted W3C standards such as the RDF Data Cube, and industry standards. We also describe how DIACHRON contributed to standards and how project partners were involved in the various working groups. This document also describes the compatibility to the OAIS standard.

Document History			
Ver.	Date	Contributor(s)	Description
0.1	03.03.2016	Jeremy Debattista (UBONN)	Table of Contents
0.2	10.03.2016	Giorgos Flouris (FORTH)	Contribution to Section 2
0.3	10.03.2016	Hasapis Panagiotis (Intrasoft)	Contribution to Section 2
0.9	20.03.2016	Jeremy Debattista (UBONN)	First Draft
1.0	30.03.2016	Jeremy Debattista and Christoph Lange (UBONN)	Final Draft

TABLE OF CONTENTS

1 INTRODUCTION

The success of the realisation of DIACHRON is partly based on the incorporation of relevant existing standards in the platform, and the DIACHRON results should be contributed back to the community in the form of additional standards on managing the evolution and preservation of web data. Adoption of standards is a challenge in itself, as the platform has to make provisions to ensure that different modules are compatible with each other during the integration and subsequent workflow. Furthermore, the DIACHRON platform coherently provides support to these standards and ensures that the modules are always up-to-date. Apart from standards available in the Semantic Web community, the DIACHRON platform is also aware of the latest state-of-the-art industry standards.

In order to achieve this goal, and to provide an interoperable solution, the project consortium had to be aware of the different standardisation activities across the different domains relevant to DIACHRON. Successful integration of these standards lowers the adoption barriers of the DIACHRON platform, whilst increase the exploitation of the project results in such communities. Moreover, apart from the identification of standards, the DIACHRON consortium partners were encouraged to participate and get involved in the various standardisation groups.

This document describes the standardisation activities undertaken in the DIACHRON Project. We describe how various aspects of the DIACHRON framework adopted W3C standards such as PROV, the RDF Data Cube, and industry standards. We also describe how DIACHRON contributed to standards and how project partners were involved in the various working groups. This document also describes the compatibility to the OAIS standard.

2 ADOPTION OF STANDARDS IN DIACHRON

2.1 ARCHIVING AND PRESERVATION STANDARDS

The work in DIACHRON was based on various standards related to knowledge representation and access. In particular, all information in a DIACHRON-conforming repository [D1.4] is represented in terms of the W3C standards RDF[1] and RDFS[2], whereas data access is mostly done using the SPARQL Query Language[3]. Operations that require updating the information in the repository (e.g., the recording of detected changes by the change detection service) use SPARQL Update Language[4], which is also a W3C recommendation. For representing multidimensional data into RDF, the W3C recommendation RDF Data Cube[5] was employed. The use of these standards ensures the easy take-up and exploitation of DIACHRON work beyond the project's lifetime.

DIACHRON is also compatible with (in fact complementary to) the OAIS standard[6]. Indeed, the OAIS Reference Model demands that each different version of a digital object should be stored using a different AIP (Archival Information Package), along with all its underlying information (metadata, authenticity, provenance, representational information etc). This would be impractical for large LOD datasets that change frequently. Our approach views the entire evolution history of a LOD dataset (i.e., all versions) as a *single digital object*, which contains the original version of the dataset, along with all its subsequent versions. This is achieved by the special structure of the DIACHRON repository, which stores all versions of a dataset in a compact and efficient manner inside the DIACHRON archive. This way, an OAIS archive can efficiently store and preserve the entire evolution history of a dataset (or large chunks of it) into one coherent AIP, without resorting to the cumbersome one-AIP-per-version approach.

2.2 INDUSTRY STANDARDS

At the implementation side, various industry standards and protocols were employed. In particular, we used industry standards for data access, including ODBC, JDBC, OLE DB and ADO.NET. Moreover, we employed standard Web and internet protocols, including HTTP, HTTPs, WebDAV, SOAP and UDDI for the implementation of DIACHRON services and for their communication. Also, JSON[7] was used as a message interchange format among various services; JSON is based on a subset of the JavaScript Programming Language standard[8], but recently evolved into a standard in its own right[9].

---

[1] RDF, W3C standard, see https://www.w3.org/RDF/
[2] RDFS, W3C recommendation, see https://www.w3.org/TR/rdf-schema/
[3] SPARQL 1.1 Query Language, W3C recommendation, see https://www.w3.org/TR/sparql11-query/
[4] SPARQL 1.1 Update Language, W3C recommendation, see https://www.w3.org/TR/sparql11-update/
[5] RDF Data Cube, W3C Recommendation, see https://www.w3.org/TR/vocab-data-cube/
[6] ISO standard, ISO-14721:2012, Space data and information transfer systems, Open archival information system (OAIS), Reference model, see http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284
[7] JavaScript Object Notation (JSON), see http://www.json.org/
[8] ECMAScript Language Specification standard, ECMA-262 standard, see http://www.ecma-international.org/publications/files/ecma-st/ECMA-262.pdf

Furthermore, we employed Web Content Accessibility Standards[10] in the implemented tool D2V in order to make the web interfaces of D2V more accessible to people with disabilities. Web content generally refers to the information in a web page or web application (like D2V), including: natural information such as text, images, and sounds, code or markup that defines structure, presentation, etc. A variety of web accessibility evaluation tools can be found under the supervision of W3C[11]. For evaluating D2V we used the Web Accessibility Checker[12] to ensure that our web application employees the WCAG 2.0 (Level AA) standard[13].

For the purposes of DIACHRON and Work Package 6, we have employed a number of standards in terms of the implementation of web services. The Hypertext Transfer Protocol (HTTP) is the application protocol that was used in to order to achieve distributed, collaborative, hypermedia information systems. HTTP is the foundation of data communication for the Web and serves as the basis of the RESTful, an architecture paradigm for the web services APIs to adhere. REST has been the standard for HTTP web services to follow. Alongside this, the message serialization has been based in JSON and JSON-LD, all being important standards, with the latter being a new W3C one. JSON is an open-standard format that uses human-readable text to transmit data objects consisting of attribute value pairs. It is the most common data format used for asynchronous browser/server communication (AJAJ). JSON-LD on the other hand, is designed around the concept of a "context" to provide additional mappings from JSON to an RDF model - the context links object properties in a JSON document to concepts in an ontology. Finally, the integration layer has employed Java Messaging Services technologies in order to provide asynchronous communication for some of the services that were developed.

2.3 DATA QUALITY

The Dataset Quality Ontology (daQ) adheres to the semantic web best practice of reuse. In fact, the daQ ontology is based on two W3C standards – RDF Data Cube[14] and the Provenance Ontology PROV-O[15]. The Data Cube vocabulary arranges *observations* (in the case of daQ [D5.2]: the values of every individual metric for every dataset under assessment) in a multi-dimensional cube. Our dimensions of interest are: metric, dataset, and time of assessment. PROV-O is used to represent the setting in which the quality assessment was carried out, including the configuration used (e.g. the quality assessment tool), to enable reproducibility and traceability. The daQ ontology is also part of the core of the Data Quality Vocabulary (described in 3.1).

---

[9] The JSON Data Interchange Format, ECMA-404 standard, see
http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf
[10] Web Accessibility Initiative (WAI)
https://www.w3.org/WAI/
[11] Web Accessibility Evaluation Tools List
https://www.w3.org/WAI/ER/tools/
[12] Achecker Web Accessibility checker
http://achecker.ca/checker/index.php
[13] Web Content Accessibility Guidelines (WCAG), Version 2.0, Level AA
https://www.w3.org/TR/WCAG20/#a
[14] https://www.w3.org/TR/vocab-data-cube/
[15] https://www.w3.org/TR/prov-o/

2.4 DATA CITATION

In order to facilitate Data Citation (i.e. accommodated temporal and provenance annotations) in DIACRHON, the PROV-O and other standards such as Dublin Core for metadata and XML Schema's *xsd:dateTime* type were used [D2.2]. Generally, these standards do not cater to specific application scenarios; for example, PROV has well-known limitations when it comes to describing scientific workflow provenance, and there are some ongoing efforts to remedy this through ontologies extending PROV such as the Wf4Ever project's Research Object ontology[16] or the DataONE project's ProvONE ontology[17]. Likewise, PROV is not necessarily enough to meet all of the needs of client applications built on DIACHRON, but if this is the case, we anticipate that client applications may be able to reuse existing ontologies (for example, for workflow provenance) or develop ad hoc extensions to cater for features not present in PROV. Therefore, in DIACHRON we extend our model with such standards, and citation meta-information is added to DIACHRONIC instances as needed.

[16] http://wf4ever.github.io/ro/

[17] http://vcvcomputing.com/provone/provone.html

3 DIACHRON CONTRIBUTION TO STANDARDS

3.1 DATA QUALITY ONTOLOGY

daQ is a light weight meta-model that enables users to create metadata for datasets related to their quality. The ontology offers the flexibility of allowing users to create their own quality schemas, making them interoperable with existing quality schemas.

The daQ ontology served as basis to the Data Quality Vocabulary[18] (DQV), a future W3C standard vocabulary, being developed by the W3C Data on the Web Best Practices[19] group. In Figure 1, we show the editors draft from February. The red box shows the overlap between the daQ ontology and the DQV.

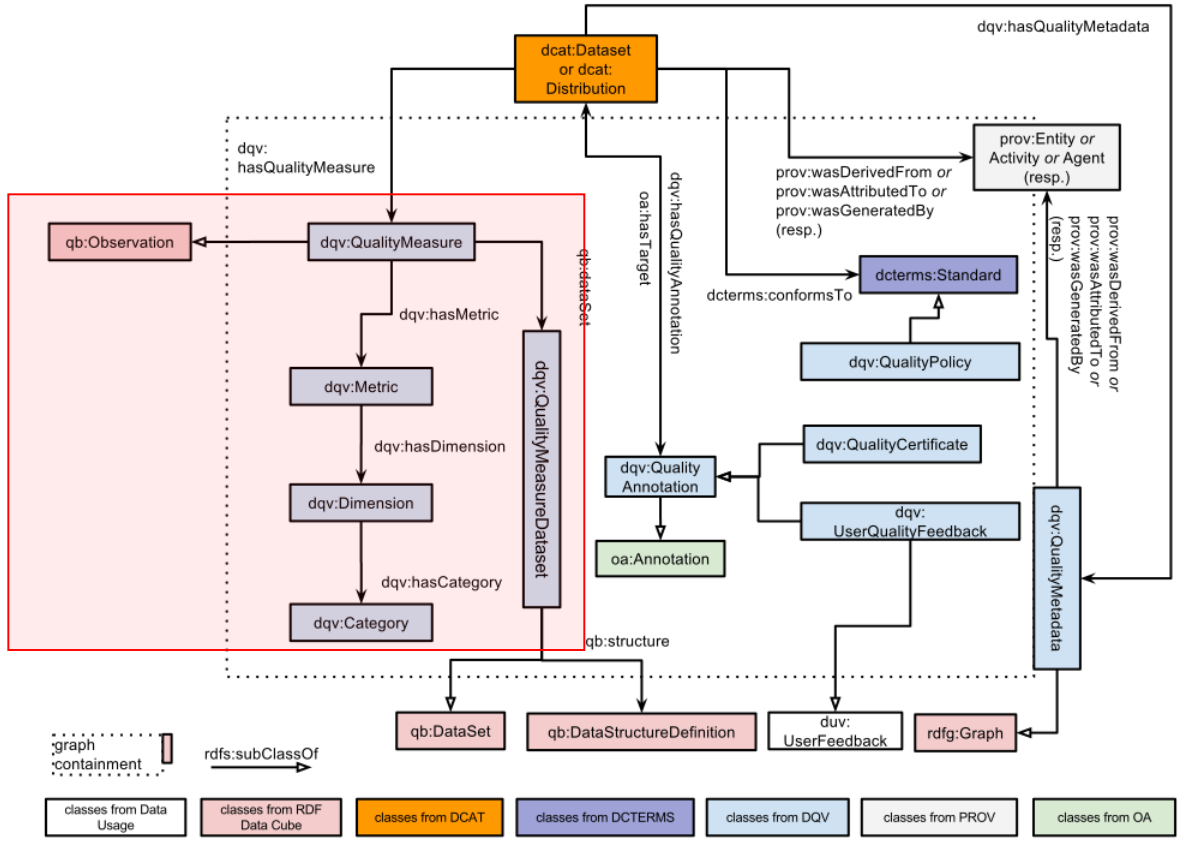

FIGURE 1 - THE DATA QUALITY VOCABULARY DRAFT AS OF 1ST FEBRUARY 2016

[18] http://www.w3.org/TR/vocab-dqv
[19] http://www.w3.org/2013/dwbp/

4 INVOLVEMENT IN STANDARDISATION ACTIVITIES

4.1 DATA ON THE WEB BEST PRACTICES W3C WORKING GROUP

The University of Bonn was heavily involved in the Data on the Web Best Practices Working Group during the duration of the DIACHRON project. The involvement included weekly conference calls with other participants, discussions in mailing lists and attending face to face meetings. The group focus was to deliver (1) a best practices document; (2) a data quality vocabulary; (3) a data usage vocabulary. University of Bonn is a key player for the output of the Data Quality Vocabulary.

4.2 PROVENANCE W3C WORKING GROUP

The University of Edinburgh was involved in the Provenance W3C incubator group. The group was involved in the development of a number of ontology recommendation and notes related to the provenance on the web. Furthermore, the University of Edinburgh was directly responsible (as an editor) of the PROV-Constraints[20] recommendation document and the note on Semantics of the PROV data model[21].

[20] https://www.w3.org/TR/2013/REC-prov-constraints-20130430/

[21] https://www.w3.org/TR/2013/NOTE-prov-sem-20130430/