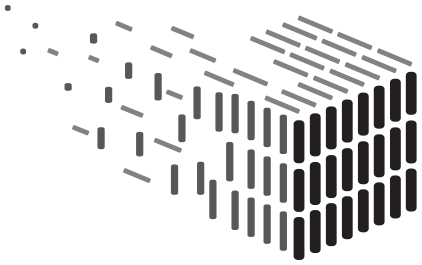

D3.3.2 Ontological Framework for a Semantic Digital Archive

DURAARK

FP7 – ICT – Digital Preservation
Grant agreement No.: 600908

Date: 2013-03-31
Version 1.0
Document id. : duraark/2013/D.3.3.1/v1.0

Grant agreement number	:	600908
Project acronym	:	DURAARK
Project full title	:	Durable Architectural Knowledge
Project's website	:	www.duraark.eu
Partners	:	LUH – Gottfried Wilhelm Leibniz Universitaet Hannover (Coordinator) [DE] UBO – Rheinische Friedrich-Wilhelms-Universitaet Bonn [DE] FhA – Fraunhofer Austria Research GmbH [AT] TUE – Technische Universiteit Eindhoven [NL] CITA – Kunstakademiets Arkitektskole [DK] LTU – Lulea Tekniska Universitet [SE] Catenda – Catenda AS [NO]
Project instrument	:	EU FP7 Collaborative Project
Project thematic priority	:	Information and Communication Technologies (ICT) Digital Preservation
Project start date	:	2013-02-01
Project duration	:	36 months
Document number	:	duraark/2014/D.3.3.1
Title of document	:	Ontological Framework for a Semantic Digital Archive
Deliverable type	:	Report
Contractual date of delivery	:	2014-01-31
Actual date of delivery	:	2014-01-31
Lead beneficiary	:	TUE
Author(s)	:	Jakob Beetz <j.beetz@tue.nl> (TUE) Michelle Lindlar <michelle.lindlar@tib.uni-hannover.de> (LUH) Stefan Dietze <dietze@l3s.de> (L3S) Ujwal Gadiraju <gadiraju@l3s.de> (L3S) Dag Field Edvardsen <dag.fjeld.edvardsen@catenda.no> (Catenda) Lars Bjørkhaug <lars.bjorkhaug@catenda.no> (Catenda)
Responsible editor(s)	:	Jakob Beetz <j.beetz@tue.nl> (TUE)
Quality assessor(s)	:	Raoul Wessel <wesselr@cs.uni-bonn.de> Martin Tamke <Martin.Tamke@kadk.dk> Östen Jonsson <osten.jonsson@ldb-centrum.se>
Approval of this deliverable	:	Stefan Dietze <dietze@L3S.de> (LUH) – Project Coordinator Marco Fisichella <fisichella@L3S.de> (LUH) – Project Manager
Distribution	:	Public
Keywords list	:	Semantic Enrichment, Ontologies, Archival, Preservation

Executive Summary

In this deliverable, a framework for a Semantic Digital Archive (SDA) for buildings and their components is suggested and described: The organizational framework for the Semantic Digital Archive as well as its methodological and technological enablers are outlined and specified. Conceptual approaches for the creation, maintenance and use of mappings between the building-specific Industry Foundation Classes (IFC) schematic and instance models with established datasets and vocabularies (e.g. the ones offered by the Linked Open Data cloud) are provided. It is shown how such mappings can be used to semantically enrich Building Information Model instances and how such enrichments can be preserved independently of the continuous evolvement of the referenced data sets. In order to track such changes in external data sets and reduce the amount of storage needed in the SDA, the concept of a Semantic Digital Observatory (SDO) is introduced and discussed. Using concrete data sets with a high relevance for the building industry, and in particular the reference vocabulary of then bSDD examples are provided that demonstrate the mechanisms introduced here. During future activities of the DURAARK project, the conceptual approaches introduced here are implemented in software prototypes as a proof of concept solution for the long-term preservation of semantically enriched Building Information Models.

Table of Contents

1 Introduction

The novel approach of the DURAARK project in comparison to earlier efforts in the domain of digital preservation of building related information is the consideration of open, self-documenting standards for Building Information Model (BIM) as well as the enrichment and correlation of architectural models with related Web data further describing its context. This approach applies to both, the building models themselves as well as their metadata used to describe the Information Packages contained in the archival system. The latter aspect helps to make extensive archives of large amounts of Information Packages searchable by additional criteria such as "buildings in the Rhineland area" or "critically debated skyscrapers".

While earlier efforts were focused on the preservation of proprietary, binary file formats such as Autodesk's DWG and DXF on a byte stream level [8][10], the DURAARK project makes distinct use of open, text-based formats from the family of the ISO 10303 standards referred to as the STandard for the Exchange of Product data (STEP). In particular, it focuses on the preservation of the Industry Foundation Classes (IFC) models along with related open specifications published and governed by the buildingSMART organization. This model has been identified as the most suitable choice for sustainable long-term archival. This model features around 650 entity classes with approx. 2000 schema-level attributes and an additional set of several thousand standardized properties that can be attached to individual entity instances. Furthermore, these properties can be conveniently extended to describe additional information specific to a particular domain, regional context or company-level data administration standards. These semantic enrichment mechanisms in fact provide a meta-modeling facility to end-users and software vendors alike while at the same time being interoperable with legacy software tools. At present however, most of this much-needed provision of vital information beyond the boundaries of the fixed schema are executed in a semantically weak and ad hoc manner. As is further detailed in section 3.1, the current practice is mostly restricted to providing key-value pairs employing only strings which are not machine-interpretable, e.g. "OK NN:+1,1m"[1]. This makes the extraction of information from mere data the harder the more distant the respective context is from the time of archival.

To overcome these limitations of string-based annotations and enrichments of engineering

[1] German abbreviation expressing that the upper edge is located at approx. 3'11" above the sea level ("Oberkante 110 cm über Normal Null")

and archival data, consensus has been reached in many research and development communities that semantic meaning has to be captured, processed and **preserved** in different ways.

To increase the interoperability among the many software systems used in the building industry without restricting the shareable information to the core model IFC, the need for extendible, domain-specific and semantically rich vocabularies has been identified time and again. Currently, the ISO 12006-based International Framework for Dictionaries (IFD) and its reference implementation buildingSMART Data Dictionary bSDD are the most widely accepted approaches to achieve this and are increasingly adopted by the industry.

With regard to Long Term Preservation (LTP), however, the use of such dynamic and networked vocabularies impose new challenges for archival: Not only the main representations of a building, such as Building Information Models, Point Clouds and other documents (not in the scope of the DURAARK project) have to be stored in Information Packages. The vocabularies and data sets referenced by these models as well as the metadata describing them have to be preserved as well. Such vocabularies and data sets are constantly evolving and mutating and hence are a "moving target" that cannot be simply referenced for later reuse without capturing the temporary state in which they have been used for the enrichment.

At present, mechanisms to link such external vocabularies to models in the building industry are mostly focused on the approach laid out in the IFD standard. The specific structure of this vocabulary as well as the technological implications rooted in the legacy of the dated STandard for the Exchange of Product data (STEP) framework would require one-of-a-kind versioning and preservation solutions. These would hardly be generalizable and applicable to other data potentially useful for the semantic enrichment of engineering models. In the DURAARK project we thus propose to apply approaches from the Semantic Web effort to address this issue. This strategy has several benefits which include:

- To harness the maturing methods and technologies developed by the large communities in the Semantic Web initiative.
- To enable the use of a wide spectrum of additional data sets including vocabularies for the semantic enrichment of engineering models
- To enable the use of distributed, networked vocabularies that can be tailored to regional, domain-specific or organizational needs without relying on a centralized,

monolithic structure as currently imposed by the bSDD.

In turn, the long-term preservation approaches for external Linked Data devised in the DURAARK project could be generalized and applied to other domains thereby increasing the contributions of the project. In the Semantic Web effort, a wide spectrum of methods, technologies and implementations to capture semantically meaningful data including vocabularies and ontologies has been devised. At its core, the Resource Description Framework (RDF) 2 allows to model information in transparent, interoperable and networked ways that enables capturing meaning in a machine-processable fashion that go beyond traditional means such as HTML, XML and proprietary database and file formats. A wide range of networked information ranging from small modelling vocabularies 3 to taxonomies, ontologies and extensive statistical data sets published as RDF datasets that together form what is referred to as the Linked Open Data (LOD) cloud. Central to this vision is to re-use and relate the data sets creating one giant graph that interconnects different datasets. To integrate the contents of the bSDD into this Linked Data cloud is an explicit sub-goal of the DURAARK project. In the DURAARK system mechanisms are devised that allow the preservation of Linked Data including vocabularies like the bSDD to complement traditional Open Archival Information System (OAIS)-compliant archival systems. The proposed approach to Long Term Preservation of Building Information Models semantically enriched with such Linked Data is based on three pillars that are described in the individual sections of this report:

Semantic Enrichment of IFC models using Linked Data captured in RDF graphs. In section 3 of this report we show how such an enrichment can be accomplished in a way that is backwards-compatible with existing legacy software tools. Our proposal is based on a few basic implementers' agreements concerning the existing standards. We demonstrate the viability of the proposed solution with exemplary datasets including an RDF version of the bSDD and other Linked Data vocabularies.

Semantic Digital Archive (SDA) . In section 4 we propose conceptual approaches for a component of the DURAARK system that enables the storage of temporal snapshots of Linked Datasets and the restitution of archived versions of semantically enriched Building Information Models (BIM's) in digital preservation life-cycles.

Semantic Digital Observatory (SDO) as a subcomponent that profiles, monitors and updates the data sets found on the Web and stored in the SDA. This is described in section 5 of this report.

2 Overview

In this section an overview of the purpose and function of the Semantic Digital Archive (SDA) and the Semantic Digital Observatory (SDO) in the preservation context of the DURAARK system is provided. In a first section a use case scenario is introduced which will be used throughout the remainder of the document to detail and demonstrate the individual components of the DURAARK framework illustrated in figure 1.

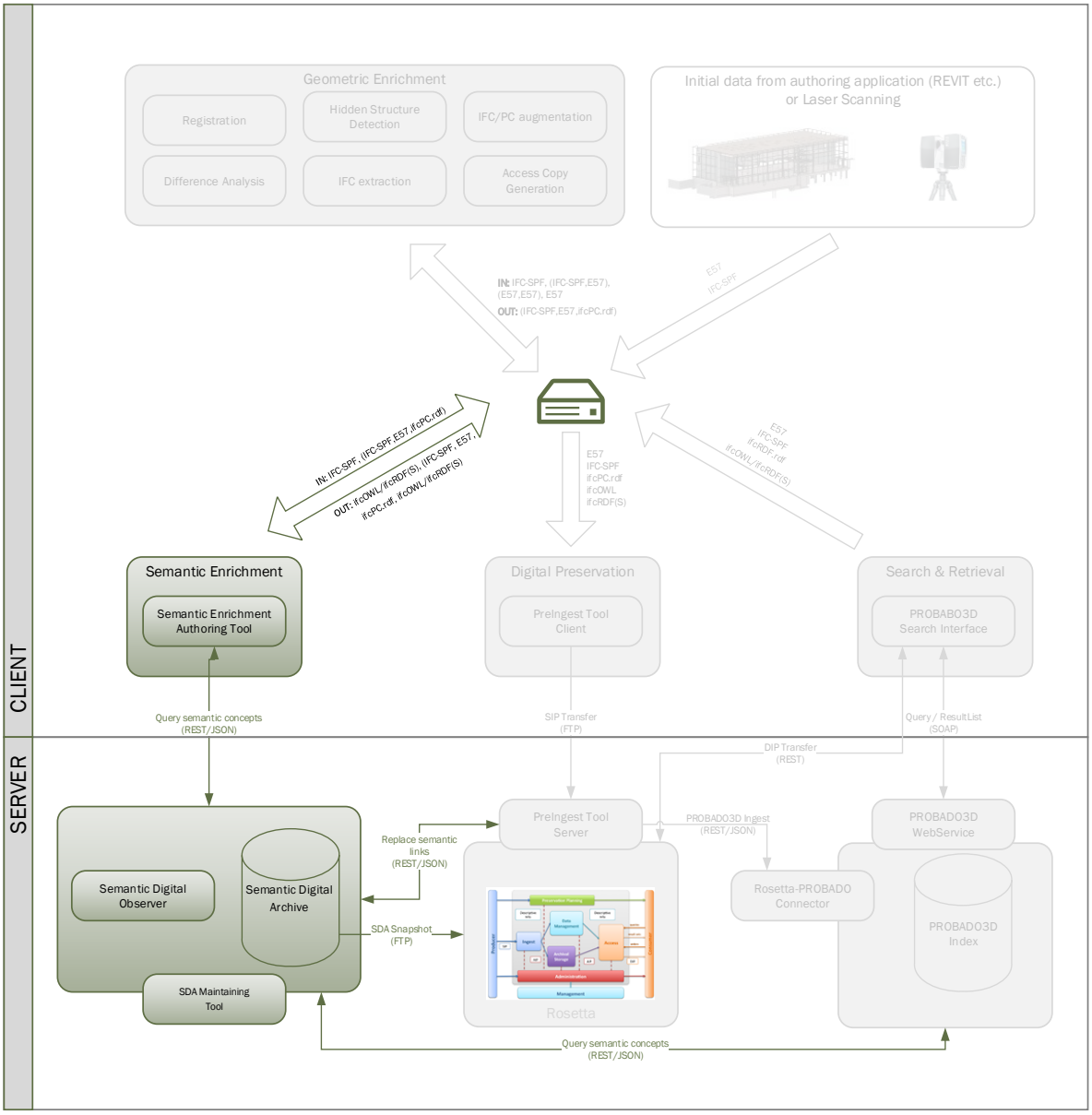

Figure 1: **Illustration of the DURAARK system components illustrating the role of the semantic enrichment, the SDA and SDO components described in this report.**

2.1 Use case Scenario

A preservation scenario documented in the related requirements documented in WP 2 is used as an example. This scenario is illustrated as a Business Process Modeling Notation BPMN diagram in figure 2. It documents the lifecycle of a building project up to its

ingestion into the Preservation System.

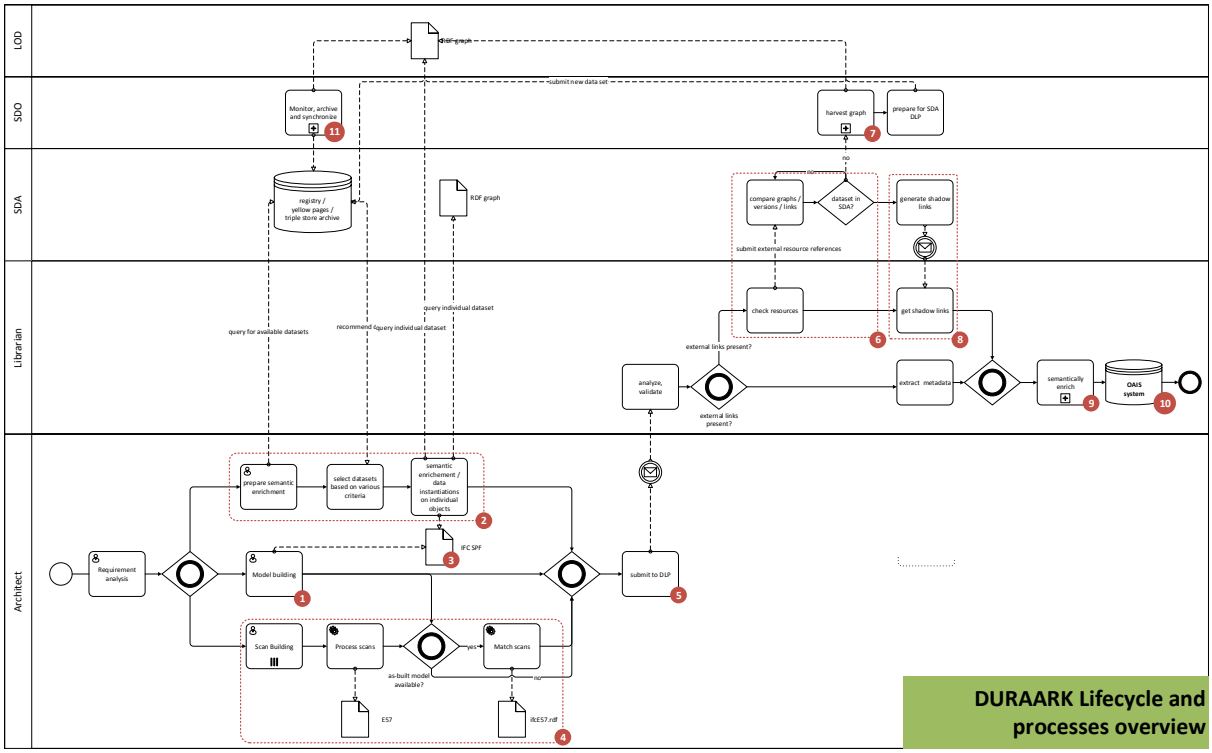

Figure 2: **Lifecycle of a building in the DURAARK system** and the intermediary sub-processes documented in this report. Please refer to section 2.1 for a full textual description. The structure of this deliverable is matched with the main parts of this illustration. Individual chapters are dedicated to the SDO, SDA and Enrichment environments

The use case scenario is divided into these individual sub-processes:

1. Model a building. *The architect or engineer creates the usual 2D and 3D CAD geometries that are semantically qualified as representing "walls", "windows" and "roofs". This standard practice can be achieved in many legacy BIM software packages. From there, the building model is exported as an IFC model.*

2. Semantically and geometrically enrich the building (components). *Where the modeling software and the IFC model fall short of detailing the semantics of the building, external datasets are employed by the architect or engineer. Examples include classifying an object as a "dormer", assigning a fire-resistance value to a door that is compliant to the respective national standard or using vendor-specific specifications of e.g. technical equipment such as ventilation units. This can be done using arbitrary datasets published as RDF graphs on the Web such as the bSDD or product*

information of individual manufacturers that will subsequently be mirrored by the SDA. As an additional service, the SDA provides a registry ('yellow pages') of all vocabularies already present in the archive and provides mappings to the individual classes of IFC model (e.g. the relevant terms of the German fire safety regulations for doors are linked to the `IfcDoor` *class)*

3. Save (semantically enriched) model to SPF. *Using a small set of implementation agreements allows to capture the enrichments in the legacy IFC model seamlessly blending into the legacy software. See section 3for details.*

4. Is an "as-built" model present? *If data from laser scans are available, the architect can optionally incorporate these additional geometrical representations into the BIM. This mash-up of representations at also allows the semantic enrichment of the point cloud data in turn. This is addressed in WPs 4 and 5 of the project*

5. Submit semantically and geometrically enriched model for archival. *The architect, engineer or librarian sends the building project to a the DURAARK system where it is further processed for long-term preservation*

6. Process the semantic enrichment contained in the IFC model: Are the Linked Datasets used for the enrichment already mirrored in the Semantic Digital Archive? *As part of the semi-automatic preprocessing tools developed in the DURAARK project, each external Linked Dataset is mirrored. Frequently re-used datasets such as the bSDD only have to be stored once and can be reused across individual Information Package in the archival system*

7. If linked dataset **is not present** in the SDA: Trigger the SDO to take a snapshot. *Where a dataset has never been used by other building models already archived in the DURAARK system, a mirror dump will be created by the SDO and stored in the SDA. Similarly, datasets already present in the SDA will be compared with the current version of the 'live' dataset for up-to-datedness.*

8. If Linked Dataset **is present** in the SDA: redirect semantic enrichments to SDA. *For each semantic enrichment encountered in the IFC model a shadow link is created that points to the archived version of the Dataset referenced by the original model. A link to the URL* `http://buildingsmart.org/bsdd#3vHdqCoT0Hsm00051Mm008`*, defining the concept with the English label "door set" that is used to lend meaning to*

a particular object in the IFC model is instead pointed to `http://duraark.eu/ex/archive/bsdd/snapshot-2014-12-01/3vHRQ8oT0Hsm00051Mm008` *which represent the state of the bSDD glossary at the time of its reference*

9 Semantically enrich metadata description of the Information Package *Using similar mechanisms described in* 2 *for the enrichment of the engineering dataset itself, a curator uses external Linked Dataset such as metadata vocabularies discussed in D3.3.1 or sentiments harvested from social media to describe the archived building*

10 Submit to OAIS *The packaged information is sent into the OAIS-compliant archival system, e.g. a Rosetta installation employed in the DURAARK proof of concept version*

11 Continuously monitor, and synchronize the mirrored vocabularies in the SDA *Triggered by a scheduler that is configured depending e.g. on the expected frequency of changes in datasets the SDO compares the mirrored datasets on a regular basis to kept the archived versions up to date*

3 Semantic Enrichment

As part of the processing tools developed for the ingestion and preservation of Building Information Models, an essential component of the DURAARK system facilitates the semantic enrichment and annotation of both content data (the BIM models preserved in the archive) and metadata (used to describe the contents of the Information Packages stored in the archive).
Semantic enrichment exploits both, expert-curate domain models and heterogeneous Web data sources, in particular Linked Data, for gradually enriching BIM/IFC models and archival descriptions with related information. Two general forms of enrichment can be distinguished:

Manual enrichment of engineering data. During the creation and modification of initial BIM/IFC models individual objects in the building assembly are enriched by architects and engineers. For example, general functional requirement specifications of a particular door set in early stages of the design ("door must be 1.01 m wide and have a fire resistance of 30 min according to the local building regulation") are gradually refined with the product specification of an individual manufacturer that has been chosen as (*"Product type A of Vendor B, catalogue number C, serial number D in configuration E3 with components X, Y, Z"*). While a number of such common requirements and product parameters can be specified using entities and facets of standardized model schemas such as the IFCs, a great deal of information is currently modeled in a formally weak and ad hoc manner. To address this, a number of structured vocabularies have been proposed in the past but have fallen short of wide adoption due to their limited exposure via standard interfaces. This includes the buildingSMART Data Dictionary bSDD. This vocabulary, which has evolved over decades[2] currently contains some 80k concepts along with approx 200k natural language names and descriptions. While currently limited to custom SOAP and REST web services, the DURAARK project exposes this information as a 5 star Linked Data Set[3] preserved as part of the SDA. A prototypical transformation of the bSDD content into an OWL ontology has been created in an initial phase of the project[4]. The experiments described in this section have been carried out

[2] see also A

[3] http://www.w3.org/DesignIssues/LinkedData.html

[4] initial experiments with transformation using less formally rigid modeling approaches such as SKOS have been carried out. These will be evaluated and discussed more in-depth in future phases of the project

using this transformed dataset. Its use and implications for the semantic enrichment approaches suggested in the DURAARK project are further discussed in sections 3.1 and 4.

Automated and manual interlinking and correlation with related Web data. As part of this step, archival information describing architectural models (BIM/IFC) are enriched with related information prevalent on the Web. Examples of such enrichment include the geo-location of a building, its history, surrounding traffic, transport and infrastructure and the usage and perception by the general public. Building on previous work on entity linking, data consolidation and correlation for digital archives, dedicated algorithms for the architectural domain are developed, tailored to detect data relating to specific geospatial areas or to specifically architecturally relevant resource types. Additionally, during the ingestion for archival which is carried out by librarians and archivists or members organizations such as municipalities, construction companies and architectural offices, other types of data sets need to be referenced. The mechanisms proposed in the DURAARK system are introduced in section 3.3.

3.1 Semantic enrichment: Current State of the Art and its limitations

As outlined in the introduction to this report and this section, the need to semantically enrich engineering data with structured vocabularies has been identified repeatedly in the building and construction industry [11, 6, 18]. A concise overview of past developments discussing the various initiatives in the building industry can be found in [19].
Current best practices and accepted approaches endorsed by the buildingSMART organization suggest the use of a single concept repository (*the* buildingSMART Data Dictionary bSDD). In recent years this approach has received most attention in the scientific and standardization communities and is currently the most likely candidate to receive support by commercial software vendors. Conceptually, the bSDD is based on the ISO 12006 series of standards, that include a theoretical framework for the principal organization of information[15] and a concrete data model to store such structures[14]. Among its many potential applications, the bSDD is intended as a dynamic extension mechanism to augment the limited semantic scope of the IFC model [8] and forms one of the three

main pillars or the buildingSMART interoperability standards[5]. In order to use the bSDD vocabulary with BIM objects captured in an IFC model, a mechanism has been proposed that uses existing facilities in the IFC schema specification to link individual vocabulary items to object instances. This proposal has been standardized in SN/TS 2489[27]. A detailed technical description of the mechanisms devised in this standard along with a concrete example of its usage is described in appendix B. A number of limitations can be identified that inhibit the use, commercial adoption and sustainable governance (including the Long Term Preservation (LTP)) of these standards. These issues include:

- The data of the underlying ISO 12006 part 3 model that captures the vocabulary is non-trivial and requires a considerable implementation effort. This effort is increased by the fact that it is specified and built on the STandard for the Exchange of Product data (STEP) technology stack which is dated and suffers from a lack of adoption among developers and poor support by software development tools.

- The bSDD vocabulary itself is contained in an information silo that can only be explored, queried and modified using a one-of-a-kind Application Programming Interface (API) of more than 50 individual function calls[6]. Although these are built on the principles of RESTful services[10][7] they make its use cumbersome and do not allow e.g. standardized ways of interlinking its contents with other vocabularies

- National building regulations and other cultural contexts, organizational information requirements and the wide spectrum of highly specialized sub-domains in the building and construction industry create the demand for highly extensible and dynamic vocabularies. However, the current bSDD approach is monolithic[8]. The limitations inherent to the 30-year old technological foundations which pre-date the wide adoption of networked information structures hinder the creation of distributed and granular vocabularies.

- Long term preservation and versioning strategies have to be specially developed and tailored to the specific characteristics of the model. This way, existing approaches

[5] http://www.buildingsmart-tech.org/specifications

[6] http://bsdd.peregrine.catenda.com/

[7] which is a considerable enhancements of earlier SOAP approaches

[8] Although the concept of individual 'contexts' has been introduced, these contexts must exist within the closed world of a single database. Additionally, these context do not cover all aspects and do not allow to e.g. contextualize the provenance of individual relations

and technologies (as e.g. further discussed in section 4.4) cannot be applied in straight-forward ways.

To address these and further issues discussed in-depth in [8, 1, 2], we are proposing to employ methods and technologies from the Semantic Web effort. To employ more rigid semantics in Building Information Modeling, several suggestions to use the Resource Description Format (RDF) and harness description logics-based modeling facilities as the Ontology Web Language (OWL) have been made in the past [4, 3, 32]. However, such transition would result a rather radical shifts of the complete stack of accepted and increasingly implemented technologies (STEP). Considering the significant investments that have been made in recent years by the sector the suggestions of such extensive shifts is very unlikely to succeed.
Instead, we are proposing a transitional solution that allows the combination of legacy STandard for the Exchange of Product data (STEP)-based IFC models with RDF data. This allows a much streamlined and simplified use of the contents of the bSDD vocabulary using standards, technologies and software tools accepted in many industries and communities beyond the boundaries of the building and construction industry. This also allows creating cross-links to other controlled vocabularies which is beneficial in many interoperability scenarios. Most importantly, the Long Term Preservation (LTP) strategies developed in the DURAARK project are applicable on a much more general level and at the same time benefit from a much more rigid and mature set of methods and technologies. The proposed gradual and backwars-compatible migration is based on two major steps that have been designed and tested in experimental prototypes in these early phases of the DURAARK project:

1. A mechanism to incorporate RDF data into legacy IFC models which is described in 3.2 and

2. The creation of an RDF version of the buildingSMART Data Dictionary bSDD vocabulary which can be stored, queried and referenced using accepted technologies and standards such as triple-/quad-store databases and SPARQL endpoints which together form the technological basis of the proposed SDA (see 4) After finalization the developed dataset will be published for public use[9].

[9]granted that the buildingSMART organization who are the offical owner of the dataset grant their permission

Additional mechanisms employing concepts from the Semantic Web will very likely be used in future which would allow the use of several, independent and cascading vocabularies. This principle mechanism is illustrate in figure 3. The technical details of this novel approach are further explained in section 3.2

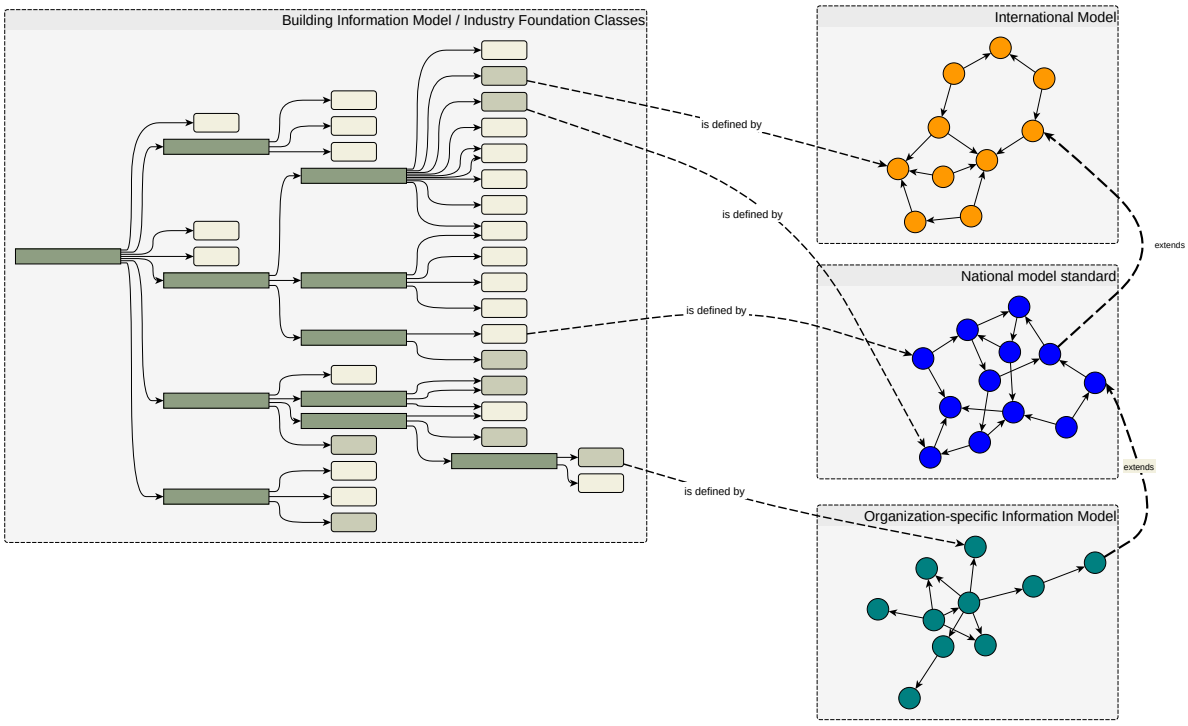

Figure 3: **Semantic Enrichment of IFC models using RDF vocabularies stemming from different sources**

A prototypical tool demonstrating the such semantic enrichment is shown in figure 4. Here, an IFC model of a quay wall structure has been enriched with additional information stemming from external vocabularies.

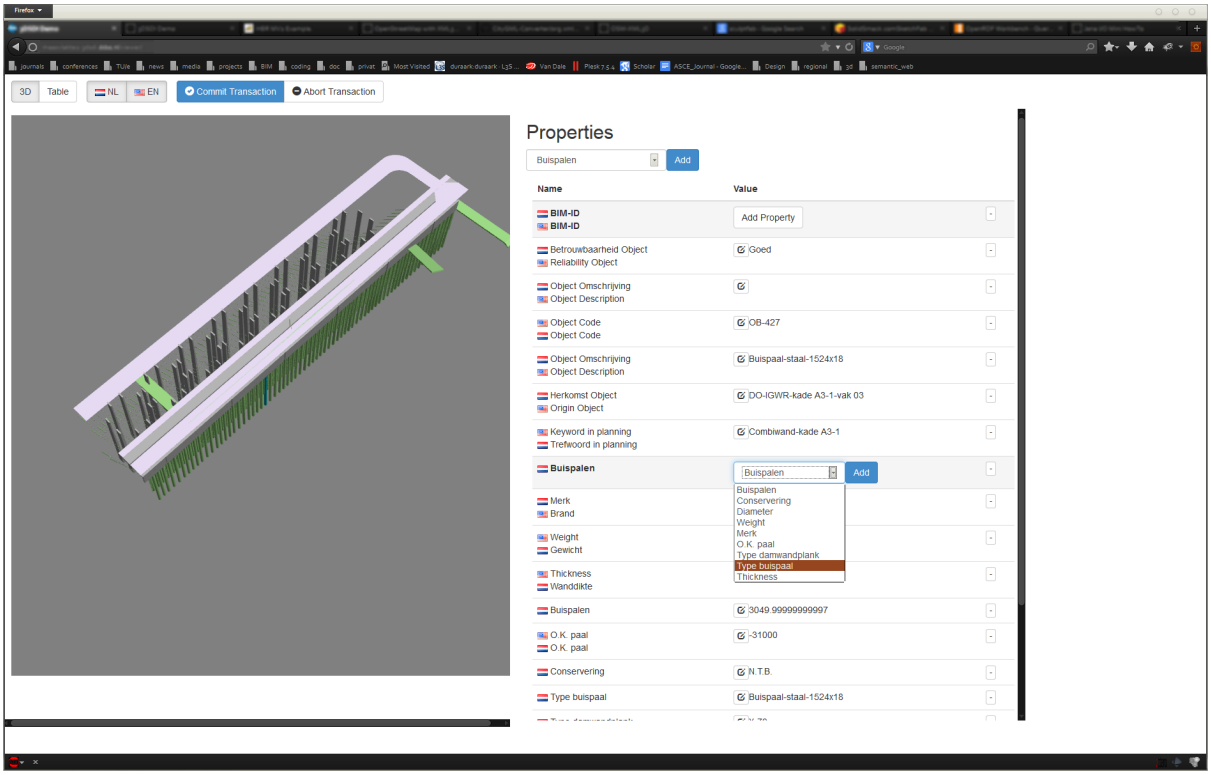

Figure 4: **Screenshot of a prototypical software tool allowing the semantic enrichment of IFC models using external RDF vocabularies. The model is provided by '3D Spatial Data Infrastructures' project funded by the "Next Generation Infrastructures" initiative**

3.2 Semantic Enrichment of IFC Data using RDF

In this section a transitional approach is suggested that allows the use of linked RDF data without breaking backwards compatibility to the STEP based IFC SPF format. The basic notion of this transitional approach is to treat each assignment of an `IfcProperty` to an `IfcProduct` using the recommended way documented in the "IFC Implementation Guide"[17] as an RDF triple statement $\langle subject, predicate, object \rangle$. A minimal example (similar to the one used for the SN/TS approach used in appendix B) is illustrated in figure 5. The excerpt of the respective minimal IFC model is provided in listing 1.

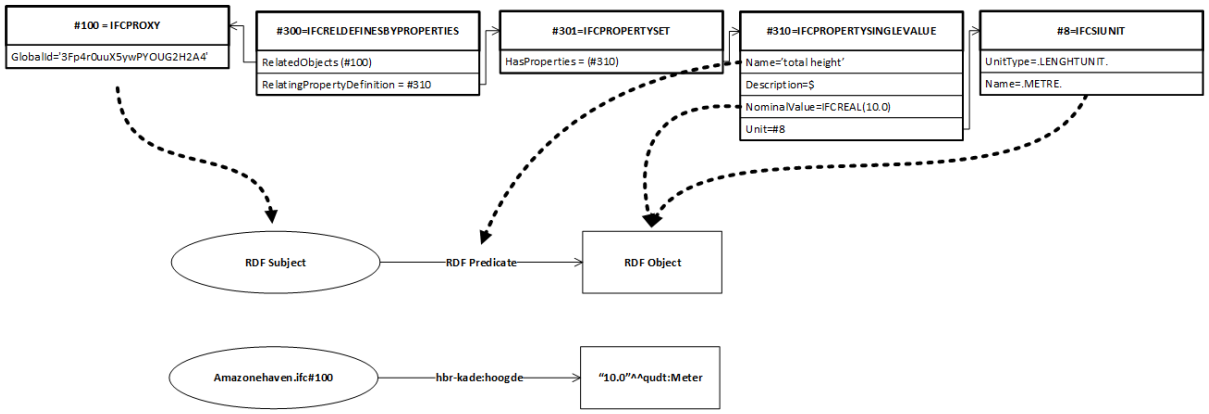

Figure 5: **Schematic illustration of semantic enrichment of IFC models using RDF and OWL vocabularies**

```
#100 = IFCPROXY('3Fp4r0uuX5ywPYOUG2H2A4', #2, 'Proxy', 'Description of Proxy', $,
    #101, #51, .PRODUCT., 'Product proxy defined externally');

#300=IFCRELDEFINESBYPROPERTIES('35YdWMmwr4rQ61AZPsifP7',#2,$,$,(#100),#301);

#301=IFCPROPERTYSET('3Fp4r0uuX5ywPYOUG2H2A5',#2,'PropertySet_With_RDF',$,(#310,#311));

#310=IFCPROPERTYSINGLEVALUE('<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>',$,
    IFCSTRING('<http://bsdd.org/vocab#3j9G00BS0Htm00025QrE$V>'),#8);

#311=IFCPROPERTYSINGLEVALUE('<http://bsdd.org/vocab#objectHeight>',$,
    IFCSTRING('\"10.00\"^^<http://qudt.org/1.1/vocab/unit#Meter>'),#8);
```

Listing 1: Partial IFC file (in the SPF format) to demonstrate the semantic enrichment of engineering data using RDF).

In this example, a building component is represented by an `IfcProxy` instance (having the local identifier `#100` in the SPF). In the IFC model, the `IfcProxy` class is *"intended to be a kind of a container for wrapping objects which are defined by associated properties, which may or may not have a geometric representation and placement in space"*[10] intended to be used, when no suitable class definition (like 'wall', 'door' or 'roof') is available for the object in the model schema.

This meaning is assigned in using the RDF predicate `rdf:type`[11] in combination with the definition of the concept "Quay Wall", having the GUID "3j9G00BS0Htm00025QrE$V" in

[10] see the formal definition at http://www.buildingsmart-tech.org/ifc/IFC2x3/TC1/html/ifckernel/lexical/ifcproxy.htm

[11] provided with the full URI in the snippet

the bSDD. This is done by using a `IfcPropertySingleValue` (local ID `#300`). Note that, while the GUID refers to the actual definition of the "Quay Wall" concept in the current bSDD the URI is (still) virtual. A small listing of the actual definition in our converted RDF representation of the quay wall concept is provided in appendix C. The (indirect) assignment via an instantiation of an objectified relationship `IfcRelDefinesByProperties` (local ID `#300`) that assigns a collection object `IfcPropertySet` is the common way that has been implemented all of the currently over hundred IFC-compliant software tools. Legacy software tools without explicit support of the proposed solution will simple treat the provided URIs as strings. In figure 6 two screenshots are provided that show how of-the-shelf legacy tools are treating the minimal test IFC file described here. In an experimental stage several common IFC tools have been tested none of which caused troubles.

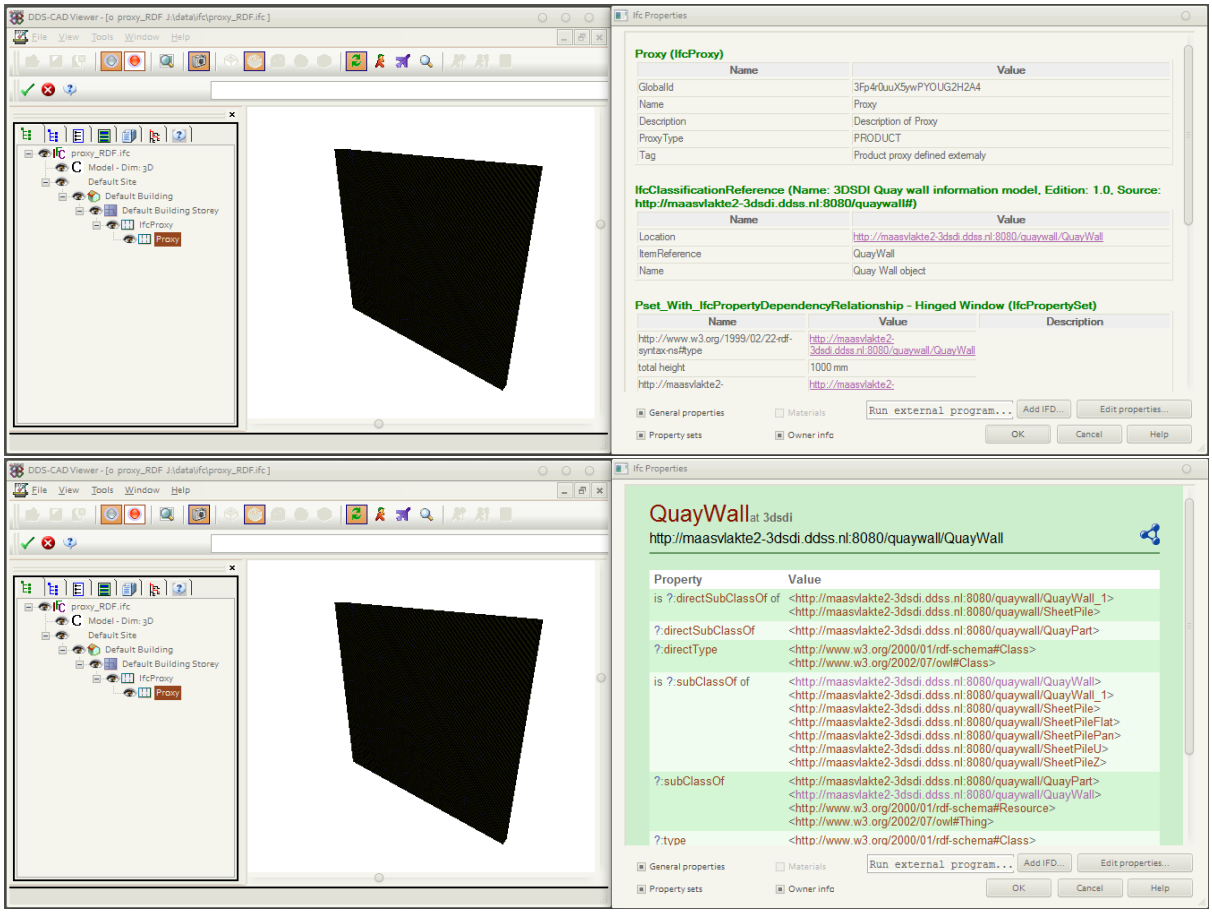

Figure 6: **Screenshots of the free legacy tool "DDS-CAD Open BIM Viewer" http://www.dds-cad.net/downloads/dds-cad-open-bim-viewer/ showing the property inspector window with the minimal RDF-enriched IFC file. The lower part shows the in-application web-browser displaying the HTML info page on the served by the SPARQL endpoint when clicking one of the URLs in the property inspector**

To fully profit from the semantic enrichment, however, software implementers have to support a very small set of agreements. In order to enable the use of RDF vocabularies for the semantic enrichment of IFC models

1. Every `IfcObject` is regarded as the *subject* of an RDF $\langle subject, predicate, object \rangle$ triple statement .

2. The `"name"` attribute of the `IfcPropertySingleValue`[12] is treated as a `rdf:Property` predicate in an RDF tripe statement.

[12] http://www.buildingsmart-tech.org/ifc/IFC2x3/TC1/html/ifcpropertyresource/lexical/ifcpropertysinglevalue.htm

3. Similarly, the `NominalValue` attribute contains the value of the property and is thus regarded as the *object* of the statement. In the case of assigning a class as the `rdf:type` of a subject (e.g. `IfcProxy`) this contains the URI of the class reference. This mechanism can also be used to assign owl:ObjectProperties. For data typed values requiring literals as simple types from the XML schema definition (xsd:float, xsd:string etc.) or more complex forms like the unit unit assignment Meter using the QUDT[13] vocabulary. in the example, an additional mapping is required. The literal can be strongly typed using the RDF mechanism for literal, e.g. prepending the literal value with a "^^<URI>" like it is used in the TURTLE[14] and N3[15] notations of RDF.

The main advantage of the approach is that only gradual changes to the technology stack have to be introduced: Even for legacy applications that have no in-built support to operate with RDF graphs and information from the Semantic Web, IFC are still syntactically correct and legible. It is merely a question of conducting an additional interpretation step to enable a wealth of additional functionality. Currently, the semantics of properties assigned in an IFC in this ad-hoc can only be interpreted using boiler plate code with string comparisons etc. The approach suggested here allows much more rigid semantics using a wide range of methods and technologies that can easily be applied and integrated into existing processes and tools.

A fully working prototypical example of this proposed mechanism is shown in figure 4 at the beginning of this section.

Our novel mechanism suggest here is currently discussed and evaluated in a number of national and international R&D projects as well as the technical bSDD panels of the buidlingSMART organization and will be presented to the larger buildingSMART community and the OpenInfra committees later in 2014. The initial and informal feedback has been very positive.

[13] http://http://qudt.org/
[14] http://www.w3.org/TeamSubmission/turtle/
[15] http://www.w3.org/DesignIssues/Notation3

3.3 Semantic Enrichment with contextual knowledge from the Web of Data

As described thoroughly in the Deliverable D3.3.1, semantic enrichment exploits existing Web Data, in particular Linked Data, in order to draw and extend meaningful insights that in turn enrich the archival data. In this section, we present a detailed example to depict how the freely available Web of data can be utilized to extract contextual knowledge for semantic enrichment. Unlike in section 3.2 which focused on examples of enriching the content data itself, the scenarios described here are used to provide additional information for the metadata attribution of the Information Packages during the pre-processing stages illustrated in 9 of figure 2.

The perception of an architectural structure at a given point in time can be archived and preserved periodically in order to gauge evolution. Historically, obtaining feedback about the perception of structures has been a challenging and costly activity. However, with the advent of the Web, a vast body of data has become available publicly. This data provides information about the perception and use of buildings, for instance through social media, and structured information about the building's features and characteristics, for instance through public Linked Data. In a work under review [16], we correlate structures with building properties described in Linked Data such as DBpedia to identify popular patterns for particular building types (airports, bridges, churches, halls, and skyscrapers). Our results show that it is feasible to mine the social and the semantic Web to create meaningful insights about popularity of architectural styles. The obtained data itself has been published through an interactive visualization in the form of a conjunct map and as public Linked Data.

We extend our approach which results in garnering perception scores of various architectural structures, by mining the Web to correlate influential factors of perception with relevant structured data.

We overcome the first hurdles towards mining patterns for well-perceived architectural structures by establishing the influential factors for different types of structures, and then generating rankings of structures based on their corresponding perception. The next challenge is to consolidate and correlate these influential factors with additional relevant information that can be extracted from the Semantic Web.

[16] Bringing Crowds to Architectural Structures - Mining the Web for Popular Architectural Patterns. Under review at *WWW 2014, Seoul, Korea.*

We exploit structured data from the DBpedia knowledge graph in order to correlate the influential factors with concrete properties and values. Table 1 depicts some of the properties that can be extracted from the DBpedia knowledge graph in order to correlate the influential factors corresponding to each structure with specific values. By doing so, we can analyse the well-received patterns for architectural structures at a finer level of granularity, i.e. in terms of tangible properties as shown in Table 1. Although we use the DBpedia knowledge graph to extract relevant data, the Web is a rich source of diverse architecture-related content and we can easily use other sources from the Web in order to make similar assertions and mine patterns for architectural structures.

Influential Factors	Airports	Bridges	Churches	Halls	Skyscrapers
History Associated, **Materials Used**, **Size**, **Level of Detail**, **Surroundings**	dbpedia-owl:runwaySurface, dbpedia-owl:runwayLength, dbpedia-owl:elevation, dbprop:cityServed, dbpedia-owl:locatedInArea, dbprop:direction [17]	dbprop:architect, dbpedia-owl:constructionMaterial, dbprop:material, dbpedia-owl:length, dbpedia-owl:width, dbpedia-owl:mainspan	dbprop:architectureStyle, dbprop:consecrationYear, dbprop:materials, dbprop:domeHeightOuter, dbprop:length, dbprop:width, dbprop:area, dbpedia-owl:location, dbprop:district	dbpedia-owl:yearOfConstruction, dbprop:built, dbprop:architect, dbprop:area, dbprop:seatingCapacity, dbpedia-owl:location	dbprop:startDate, dbprop:completionDate, dbpedia-owl:architect

Table 1: Some DBpedia properties that can be used to materialize corresponding Influential Factors.

[17] In Table 1 dbprop:direction, direction is one of north, south, east, west, northeast, northwest, southeast, or southwest

4 SDA - Semantic Digital Archive

In this section a description of the Semantic Digital Archive is provided which is an integral component of the DURAARK system.

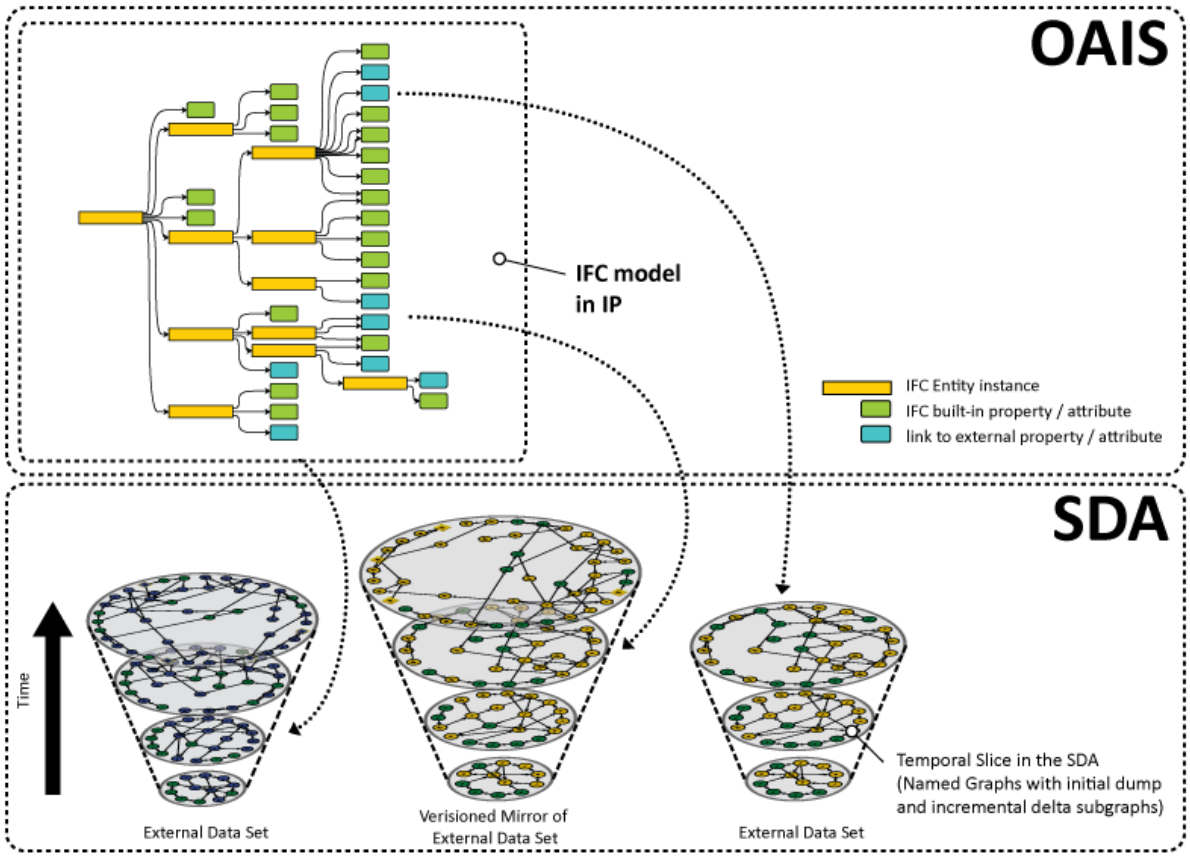

Figure 7: **Schematic illustration of creating links between individual temporal slices of mirrored vocabularies between the main OAIS-compliant DURAARK archive and the SDA**

4.1 Scope of the SDA

The Semantic Digital Archive (SDA) is the central knowledge base for structured metadata and contextual knowledge about buildings and built structures in the DURAARK system. As such, the main purpose of the SDA is to provide an accessible endpoint which allows queries on architecturally relevant knowledge and building metadata. The SDA will provide a snapshot of most relevant data to facilitate the use cases identified in earlier deliverables and support queries from DURAARK stakeholders, such as archivists, architects and

urban planners. In order to allow efficient and scalable queries, the SDA provides only a subset of metadata, for instance, covering only a restricted time period, and will regularly archive less relevant and outdated data into the DURAARK preservation system.
As illustrated in Figure 7, the SDA is populated through the enrichment components described in Sections 3 and 4 of the Deliverable D3.3.1 [18], which gradually enrich raw building data with contextual metadata. Where applicable, the SDA has to provide views of parts of the considered external datasets, requiring the preservation of related subgraphs from, for instance, DBpedia, Geonames, bSDD or other related datasets at the time of enrichment. Archiving strategies will be defined based on knowledge in the SDO about the nature and dynamics of existing datasets.
The SDA is implemented through a RDF data store, currently realised through Virtuoso[19], a widely established NOSQL storage solution, which provides a public SPARQL endpoint and capabilities for dereferencing URIs in the DURAARK dataset. The schema and vocabularies used for populating the store are described in the deliverable D3.3.1, while example instances and queries are described below.

4.2 Enriched Building Model Archives

To facilitate data reuse and take-up, one has to observe and understand the nature of existing Web datasets and their evolution over time. By extending the DURAARK schema with external vocabularies and relevant metadata from related datasets, we can gain access to additional information. For example, we can extend our schema by using the DBPedia ontology. Consequently, by querying the DBPedia graph, we can access additional information corresponding to a building. For instance, the architect of the building. The Figure 8 depicts the `dbpedia-owl:significantBuildingof` property, which can be used to access the metadata pertaining to the architect of the structure (in this example, the Empire State Building).
In addition, we have created a dataset of architectural structures by exploiting and extending existing schemas and vocabularies. We publish our dataset abridged with normalized popularity scores of these structures, harnessed using methods dependent on mining the Web of data, in the form of Linked Data by following the Linked Data principles. The knowledge base thus created, can be accessed and queried using the following

[18] D3.3.1-Metadata schema extension for archival systems.
[19] http://virtuoso.openlinksw.com/

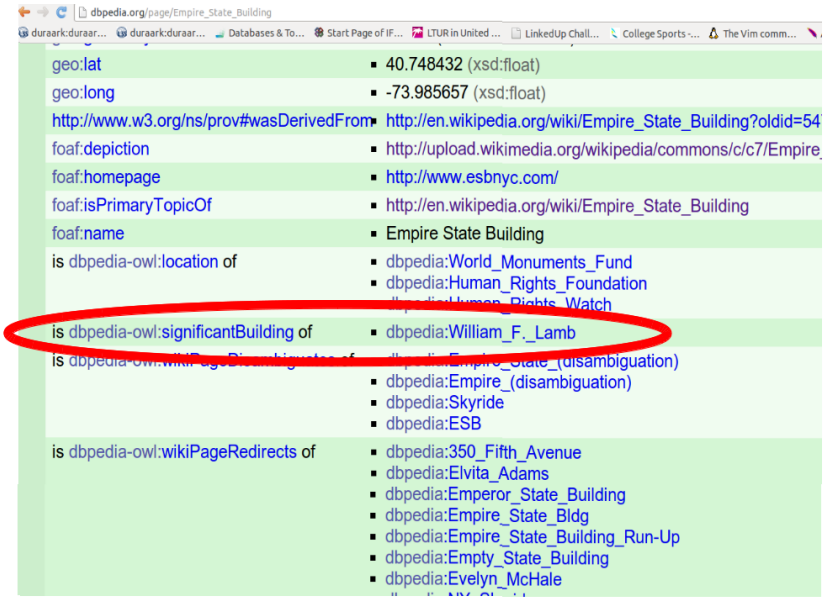

Figure 8: Accessing additional information for archival through schema extension.

SPARQL endpoint in the SDA: `http://meco.l3s.uni-hannover.de:8829/sparql`. An example query is presented in the listing 2 below. The query retrieves all the structures that have a positive perception score that is greater than 0.7 and a perception of joy that is greater than 0.2 (on a linear scale ranging between 0-1). We can obtain useful information through such intuitive queries.

```
SELECT ?structure WHERE {
?structure duraark:hasPerception_Positive ?score1.
?structure duraark:hasPerception_Joy ?score2.
FILTER ((xsd:double(?score1)>0.7) && (xsd:double(?score2)>0.2))}
```

Listing 2: Example query that demonstrates enriched building data that can be exploited.

4.3 Preservation of External (Linked) Data

Given the evolving nature of Web data, enhancement of data within the SDA has to consider the archiving of (parts of) the external data used, as a part of the enrichment process.

The graph-based and distributed nature of Linked Data has serious implications for enriching digital archives with references to external datasets. While distributed datasets

(schemas, vocabularies and actual data) evolve continuously, these changes have to be reflected in the archival and preservation strategy. This joint and simultaneous consideration of semantic enrichment and preservation aspects is usually under-reflected in archival efforts and needs to be tackled in an integrated fashion.
Generally, while within the LD graph, in theory all datasets (and RDF statements) are connected, LD archiving strategies are increasingly complex and have to identify a suitable balance between correctness/completeness on the one hand and scalability on the other. These decisions are highly dependent on the domain and characteristics of each individual dataset, as each poses different requirements and presents varying challenges with regards to the preservation strategies. For instance, datasets, differ strongly with respect to the dynamics with which they evolve, that is, the frequency of changes to the dataset. For instance, there might be fairly static datasets where changes occur only under exceptional circumstances (for instance, 2008 Road Traffic Collisions in Northern Ireland from data.gov.uk) while on the other hand, other datasets are meant to change highly frequently (for instance, Twitter feeds or Highways Agency Live Traffic Data). For the majority of datasets, changes occur moderately frequently (i.e. on a daily, weekly, monthly or annual basis) as is the case for datasets like BauDataWeb[20] or DBpedia . Depending on the specific requirements, nature and dynamics of individual datasets, we are exploring Web data preservation strategies, including the following:

1. Non-recurring capture of URI references to external entities as is common practice within the LD community.

2. Non-recurring archival of sub-graphs or the entire graph of the external dataset.

3. Periodic crawling and archiving of external datasets.

In order to inform the DURAARK preservation strategies addressed in Deliverable D6.6.1[21], and to provide efficient and scalable archiving techniques, structured knowledge about relevant dataset is required, for instance, to provide information about their endpoints, size, relevance, and in particular, their dynamics (as a whole or of sub-graphs).

[20] http://semantic.eurobau.com/
[21] Current state of 3D object digital preservation and gap-analysis report.

4.4 Versioning of evolving data sets

A major requirement for the consistent preservation in the DURAARK system is the ability to preserve and reconstruct the Linked Data sets used in the content or metadata records. As described in earlier sections, one of the main purposes of the SDA is, to capture specific states of arbitrary Linked Datasets (RDF graphs) and to reproduce the snapshots of historic states when the original has evolved and mutated. In order to prevent the inefficient pollution of the SDA through the parallel and redundant storage of largely congruent datasets, we propose the application of an approach based on changesets and Named Graphs. In this section, we are describing the principles of this approach.

Related Work Versioning continuous evolution of data published on the Web has been identified as a critical issue early on in Knowledge Engineering and Semantic Web communities [29, 24]. Previous work from the field of database and XML schema evolutions addressed these problems mainly in closed-world scenarios where schema evolution affected individual systems and their immediate context. Numerous work on ontology evolution [23, 22] and RDF evolution [5] have been contributed since and have led to dedicated versioning systems for individual graphs often adapting approaches from general purpose text-versioning systems such as SVN and GIT[34]. Approaches of version management and provenance data in dedicated management systems [28, 23, 12, 25] are not feasible for the SDA purposes, where arbitrary RDF datasets captured from external sources have to be treated as black boxes. This requires to track the changes of the datasets on the smallest level of granularity. For RDF graphs these are the individual $\langle subject, predicate, object \rangle$ triples that form the graphs. A wide range of possible approaches from different angles have been devised in recent years. Excellent overviews discussing their efficiency and successfulness are found in [21, 13].

DURAARK approach For the DURAARK SDA we are combining earlier approaches based on the notion of 'changesets' that are adapted from 'diff' and 'patch' principles[20] found in generic text-based versioning systems to RDF[5, 33, 16]. Instead of using reified `rdf:Statement`s however, we combine this with Named Graphs, which are becoming an integral part of the RDF 1.1 framework[22]. In essence, Named Graphs allow the clustering of RDF triples into contexts. These can be efficiently stored in dedicated databases that

[22] http://www.w3.org/TR/rdf11-concepts/

use quads of the form $\langle subject, predicate, object, context \rangle$ instead of triples to store RDF data [23].

The basic principle is to take an initial snapshot of an RDF dataset which is stored into a first Named Graph context G_{t_0}. When a modification of the dataset at its origin is measured upon a later visit, the difference is measured as a Delta and stored in a changeset $S_\Delta = G_{t_0} - G_{t_1}$. This changeset contains the sets of the *added* A_{t_1} and *removed* R_{t_1} triples grouped together into individual Named Graphs. The changesets themselves are attributed with additional metadata such as a timestamp. By combining the Named Graph sets of the initial snapshot with incremental changes, all earlier temporal states of the dataset can be reconstructed by

$$G_{t_n} = G_{t_0} \cup \Sigma(A_{t_1}, \ldots, A_{t_n}) - \Sigma(R_{t_1}, \ldots, R_{t_n}) \tag{1}$$

Figure 9 illustrates this using an example in which an initial triple taken from the bSDD vocabulary is changed: The string Literal "Türheinheit" that is assigned as a German name to the concept "Door set" via the `rdfs:label` relation is changed into "Tür". The difference that is measured can be captured as the removal of the triple $R_{t_1} = \langle$Concept,rdfs:label,"Türeinheit"$\rangle$ and the addition of the triple $A_{t_1} = \langle$Concept,rdfs:label,"Tür"$\rangle$. A similar change from "Tür" to "Tür mit Zarge" is detected at a later time. Using the operation in equation 1, every state of the dataset can be restored later.

[23] which can be regarded as low-level Reification

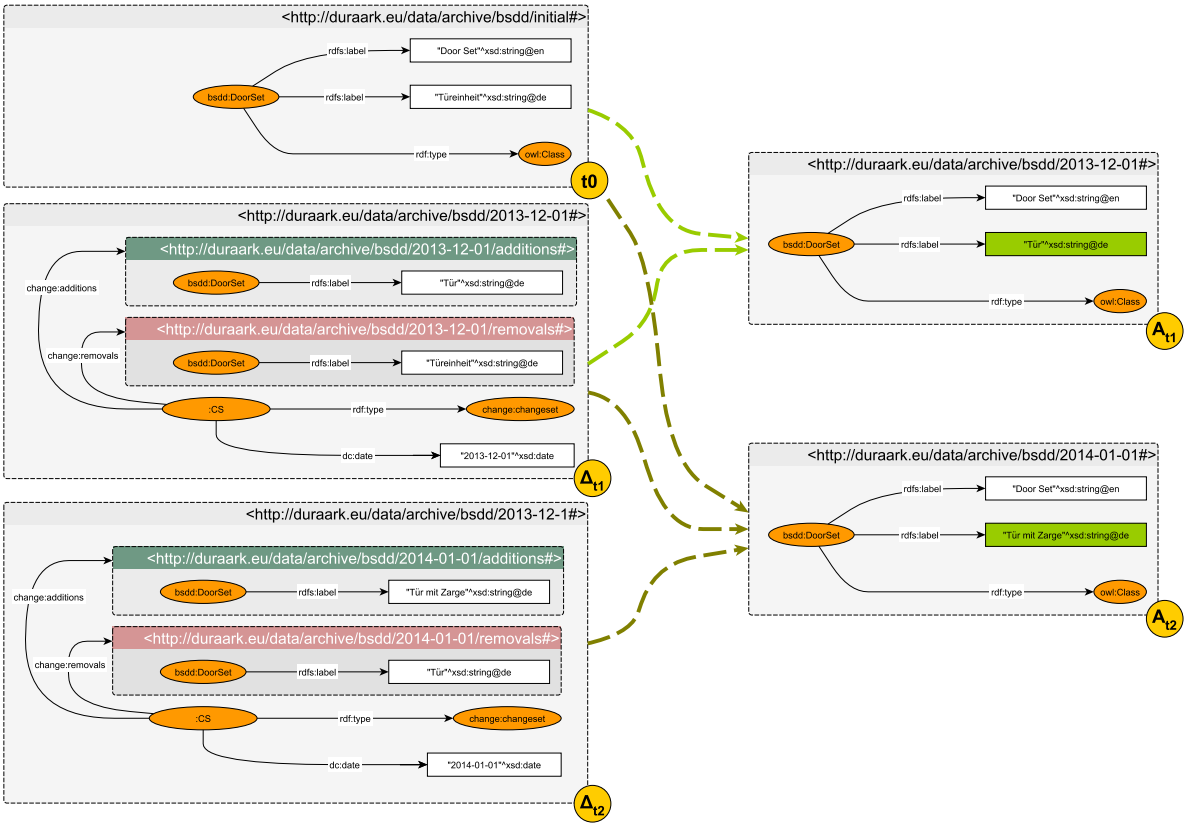

Figure 9: **Schematic overview of versioning arbitrary RDF data using Named Graphs.**

Preliminary experiments with prototypical implementations The necessary set operations implemented with real-world datasets using the SPARQL 1.1[24] language that allows the efficient computation of unions and differences among named graphs. A minimal example capturing the example described above and illustrated in figure 9 can be found in listing 7 in appendix D. Here, the original (partial) graph of the bSDD and the individual changesets containing the deltas with additions and removals are organized in separate Named Graphs using the TriG syntax[25] which is supported by some quad store implementations such as Sesame[26]. To model this, a slightly adapted version of the ChangeSet vocabulary[27] has been created, that allows the inclusion of arbitrary URIs

[24] http://www.w3.org/TR/sparql11-query/

[25] an N3 / TRUTLE variant includig the support of Named Graphs, http://wifo5-03.informatik.uni-mannheim.de/bizer/trig/

[26] http://www.openrdf.org/

[27] http://vocab.org/changeset/schema.html

(identifying the Named Graphs) instead of `rdf:Statement` for reification like in the original vocabulary. Using two subsequent SPARQL queries, a particular state can be retrieved from this repository.
At first, a list of Named Graph URIs containing the modifications before or on December 1st 2013 are retrieved in listing 3:

```
SELECT DISTINCT *
WHERE
{
    ?cs a duraark:ChangeSet .
    ?cs duraark_cs:createdDate ?date .
    ?cs duraark_cs:additions ?additions .
    ?cs duraark_cs:removals ?removals .

    FILTER (?date, >= "2013-12-01"^^xsd:date)
}
```

Listing 3: SPARQL query to retrieve all Named Graphs containing changesets of modifications made to the bSDD vocabulary on or before 2013-12-01. Prefixes omitted

Subsequently, these named graphs can be used to generate a query that retrieves all `rdfs:labels` from the `UNIION` of the initial graphs plus the additions from the changesets (done by the `FROM` keyword selecting individual graphs) without the removed triples (done with the `MINUS` SPARQL operation or a `FILTER`). This is shown in listing 4:

```
SELECT DISTINCT *
FROM <http://duraark.eu/ex/archive/bsdd/initial>
FROM <http://duraark.eu/ex/archive/bsdd/changeset/1/additions>
WHERE
{
    ?s rdfs:label ?o .

    MINUS{
        GRAPH <http://duraark.eu/ex/archive/bsdd/changeset/1/deletions>
        { ?s ?p ?o .}
    }.

}
```

Listing 4: SPARQL query to retrieve the temporal state as of 2013-12-01 of the partial bSDD vocabulary provided in appendix D, that will include the change to "Tür", but not to "Tür mit Zarge" that has been made later on 2014-01-01.

Other experiments, where these separate queries are combined much more elegantly into

a single one using dynamic variable bindings of named graphs have also been successfully tested on some triple store implementations. However, at the current time, the behaviour of queries of several parallel Named Graphs in one repository is underspecified in the the SPARQL 1.1 specification and has been implemented differently in available triple stores. Further experiments will be conducted during later implementation phases of the DURAARK project.

4.5 SDA - OAIS connection

The use of semantic registries to aid the preservation process has been suggest before, e.g., the University of Southampton's P2 Registry which aims to aid the format risk management process making use of available semantic web and Web2.0 sources[31]. While the question of how such a registry can be successfully facilitated in the digital preservation process is a central question, the question of how the registry data itself is being archived is usually not addressed.

In the DURAARK architecture, the SDA is seen as a registry/repository of relevant datasets which have been harvested from the web. It is to be used by experts from varying domains, e.g., architects and librarians. Looking at archiving practises for other repositories, certification procesdures such as the DINI (Deutsche Initiative für Netzwerkinformation e.V.)[28] Certificate for Document and Publication Services exist. While long-term availability is one of the criteria to be met as part of the this certification process, DINI explicitly points out that document and publication repositories are not necessarily trustworthy long term archival systems. According to the DINI certificate requirements, the repository needs to guarantee the availability of the objects and their respective metadata for a minimum of five years, while long-term availability can be ensured in cooperation with an archiving institution. Additionally, a few recommendations are given to the repository with the aim of ensuring archival criteria from the start, thus aiding the hand-over to a long-term archive. These recommendations include a regulated deletion process, the use of open file formats and avoiding technical protection measures such as Digital Rights Management (DRM) or password protection [26].

In the DURAARK system landscape, the SDA acts like such an intermediate repository, which stores the data for a certain interval but does not facilitate full digital preservation support. As the long-term availability of the SDA data is nevertheless important to support

[28] http://www.dini.de/

the audit trail and provenance and to furthermore allow for historical interpretations of the data, snapshots from the SDA will be passed to the OAIS compliant digital preservation system for full lifecycle support.

5 SDO - Semantic Digital Observatory

This Section elucidates the purpose and role of the Semantic Digital Observatory in the DURAARK system.

5.1 Overview

In order to enable the discovery and retrieval of suitable datasets and to identify dedicated and most efficient preservation strategies for each relevant dataset, we need to provide structured metadata about available datasets. This includes in particular preservation-related information, for instance about the temporal and geographic coverage of a dataset, the estimated update frequency or the represented types and topics. For example, whether the data contains building-related policy information or traffic or environmental data. For this purpose we are currently in the process of establishing dedicated data curation and profiling strategies for architecturally relevant Web data. Dataset curation and preservation follows a two-fold strategy:

- Semi-automated curation and preservation of distributed Web data.
- Expert-based curation and preservation of core vocabularies .

While there exists a wealth of relevant Web datasets, particularly Linked Data, providing useful data of relevance to the architectural field, the metadata about available datasets is very sparse.
Considering LD and Open Data in general, the main registry of available datasets is the DataHub[29]. It currently contains over 6000 open datasets and as part of the Linked Open Data group, over 337 datasets. However, while the range of data is broad, covering information about building-related policies and legislation, geodata or traffic statistics, finding and retrieving useful datasets is challenging and costly. This is due to the lack of reliable and descriptive metadata about content, provenance, availability or data types contained in distributed datasets. Thus previous knowledge of the data or costly investigations to judge the usefulness of external datasets are required. In addition, while distributed datasets evolve over time, capturing the temporal evolution of distributed datasets is crucial but not yet common practice. We currently conduct a number of data curation activities, aimed at assessing, cataloging, annotating and profiling all sorts of

[29] http://datahub.io/

Web data of relevance to the architectural domain (independent of their original intention) where the overall vision entails the creation of (a). a well-described structured catalog of datasets, and (b). an architectural knowledge graph which enables architects, urban planners or archivists to explore all forms of suitable Web data and content captured in our SDA. This work covers several areas:

- Data cataloging on the DataHub: similar to the approach followed by the Linked Open Data community effort, a dedicated group ("linked-building-data") has been set up (though not yet populated) to collect datasets of relevance to the architectural field. While the DataHub is based on CKAN , our group can be queried through the CKAN API, allowing further processing.
- Automated data assessment, profiling and annotation: while existing dataset annotations often do not facilitate a comprehensive understanding of the underlying data, we aim at creating a structured (RDF-based) catalog of architectural-related datasets.
- Gaining new insights and understanding about the nature, coherence, quality, coverage and architectural relevance of existing datasets.
- Automatically obtaining annotations and tags of existing datasets towards a more descriptive dataset catalog.
- Improving coherence and alignment (syntactic and semantic) of existing datasets towards a unified knowledge graph.

As part of such activities, we are currently in the process of generating a structured dataset catalog, which adopts VoID for the description, cataloging and annotation of relevant datasets. Schema (type and property) mappings facilitate an easier exploration of data across dataset boundaries. This work builds on our efforts in [9], yet we aim to not only provide metadata about the dynamics of datasets but also additional metadata about topic, spatial or temporal coverage of the data itself. Automated data assessment exploits a range of techniques, such as Named Entity Recognition (NER) techniques together with reference graphs (such as DBpedia) as background knowledge for classifying and profiling datasets, for instance, to automatically detect the geographical and temporal coverage of a dataset or the nature of the content, or whether it describes traffic statistics for the Greater London area, or energy efficiency policies for Germany.

As described in Section 3, different preservation strategies are considered for each dataset, depending on the dynamics, frequency, and size of updates. While each strategy requires knowledge about the datasets to interact with, like for instance the URI of their SPARQL endpoints, our VoID-based "Linked Building Data" catalogue will provide the basis for realizing such individual preservation strategies and will be enriched with preservation-related metadata (for example, about the update procedures and evolution of each dataset).
The methods and strategies described in the following sections form the building blocks of the Semantic Digital Observatory (SDO).

5.2 Profiling Web Datasets

As described above, profiling of Web data has to consider a number of characteristics and features, such as their dynamics, the covered resource types or topics, the quality or interdependencies of a dataset. Our initial work has concentrated on creating structured profiles of topics and types covered by Web datasets (see [9]).
While datasets are highly heterogeneous with respect to represented resource types, currentness, quality or topic coverage, only brief and insufficient structured information about datasets are available. In the case of DataHub, only simple tags, few structured metadata about the size, endpoints or used schemas and brief textual descriptions are available. This causes significant problems for data consumers to identify useful and trust-worthy data for different scenarios.
Nevertheless, earlier works address related issues [7, 30], such as schema alignment and extraction of shared resource annotations across datasets. However, they do not yet facilitate the extraction of reliable dataset metadata with respect to represented topics. In order to address these limitations, we present an approach that automatically and incrementally indexes datasets by interlinking and annotating arbitrary datasets with relevant topics in the form of DBpedia entities and categories. By incrementally computing topic relevance scores for individual datasets, we gradually create a knowledge base of dataset meta-information. To improve scalability the process exploits representative sample sets of resources. Moreover, to ensure high annotation accuracy a semi-automated evaluation approach is proposed.
Our dataset profiling platform automatically extracts top-ranked topic annotations (DBpedia categories) and captures these together with a relevance score for each dataset

description. All dataset descriptions are captured using a combination of the VoID[30] and VoL schemas[31]. Figure 10 depicts an example of a VoID-based dataset description and its baseline properties (without dataset profile information). Note that this is an example of how we envisage capturing descriptions of datasets related to DURAARK.

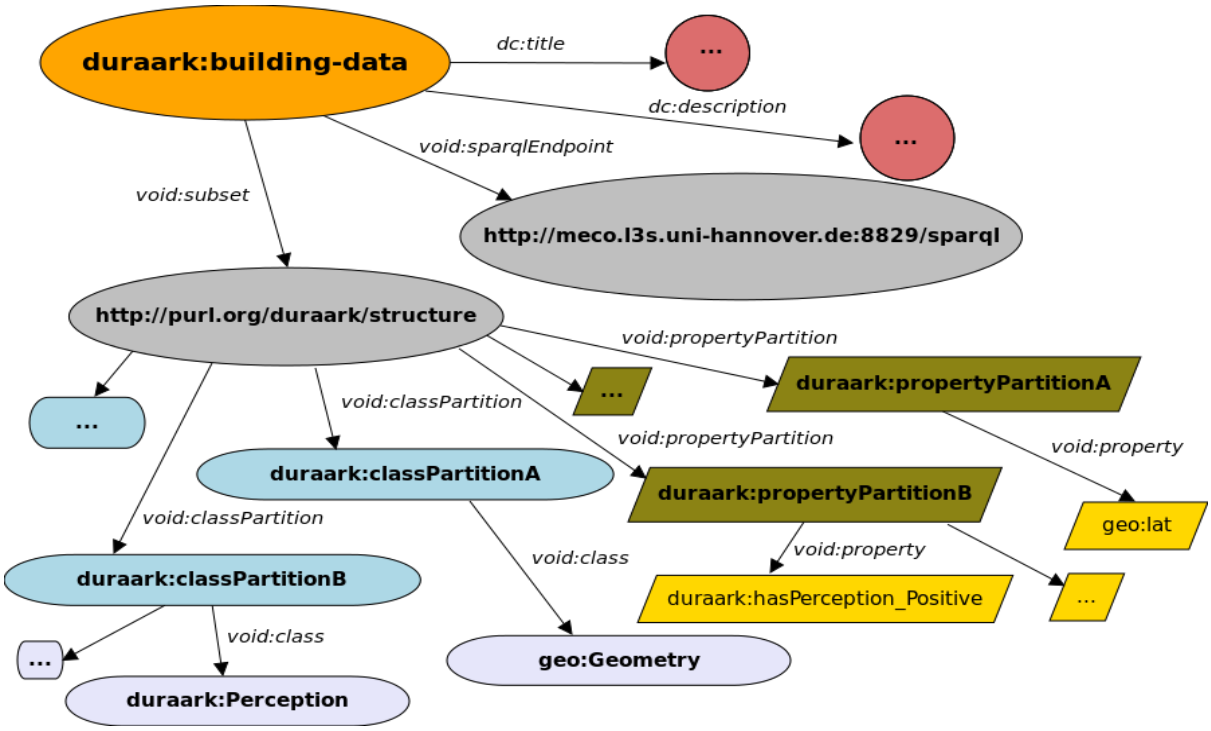

Figure 10: VoID example

Based on this, we exploit an initial prototype that has been developed in LinkedUp[32] and further refined and adopted in DURAARK. This prototype provides an exploratory means to browse and search through existing datasets in the entire Linked Open Data (LOD) Cloud according to the topics which are covered. By deploying entity recognition, sampling and ranking techniques, the prototype allows to find datasets providing data for a given set of topics or to discover datasets covering similar fields. The demo showcasing this prototype enables an exploratory search through the generated dataset profiles. Currently, the top-200 topics for each dataset profile are shown. A screenshot of the profile explorer interface is presented in the Figure 12. The generation of dataset profiles is discussed in more detail in Section **??**. An example query to get all profiles belonging to a certain

[30] http://www.w3.org/TR/void
[31] http://www.purl.org/vol/ns
[32] http://linkedup-project.eu/

category, for instance 'Architectural Styles' is presented in Figure 11. In this way, by exploiting the DBPedia category graph, it is possible to extract specific datasets relevant to DURAARK.

```
SELECT ?dataset ?link ?score ?entity ?resource WHERE {
?dataset a void:Linkset.
?dataset vol:hasLink ?link.
?link vol:linksResource
<http://dbpedia.org/resource/Category:Architectural_styles>.
?link vol:hasScore ?score.
?link vol:derivedFrom ?entity.
?link_1 vol:linksResource ?entity.
?dataset vol:hasLink ?link_1.
?link_1 vol:derivedFrom ?resource
}
```

Figure 11: Example SPARQL query for obtaining relevant dataset profiles.

By employing the dataset profile explorer, we can find relevant datasets that exist in the LOD cloud, which can in turn help in semantically enriching the archival data.

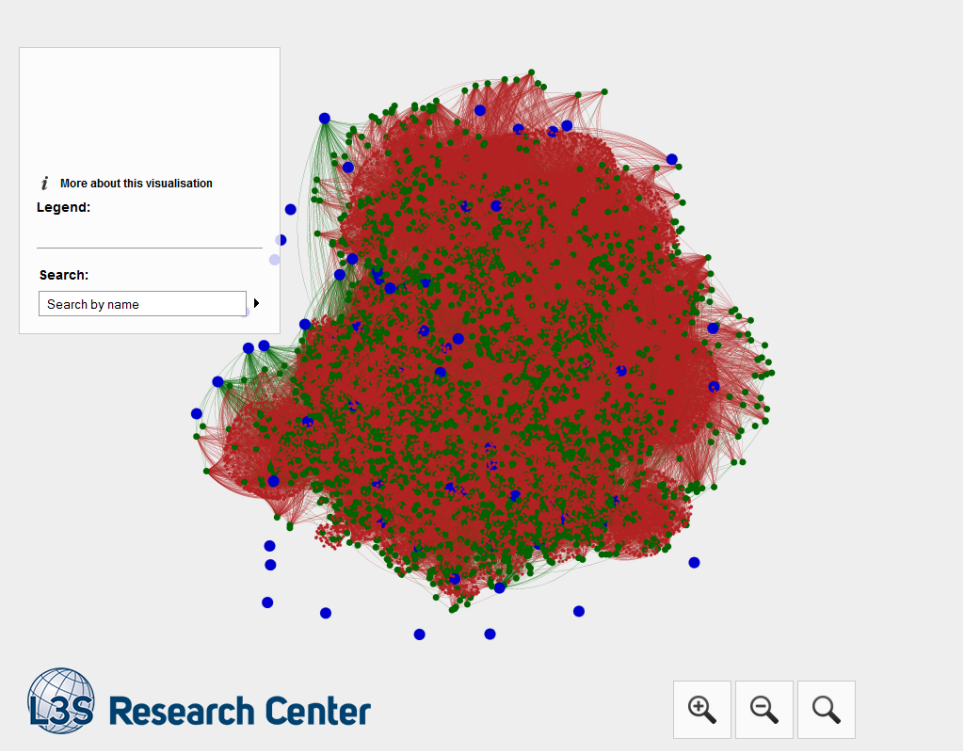

Figure 12: Screenshot of the Profile Explorer depicting our dataset profiles in the LOD cloud.

A screenshot of the profile explorer interface is presented in the Figure 12. One can select any particular dataset by zooming into the visualization, search or filter for datasets relating to their interest or information need (in our case, to find datasets relevant to DURAARK). An example query to get all profiles belonging to a certain category, for instance 'Architectural Styles' is presented in Figure 11. In this way, by exploiting the DBpedia category graph, it is possible to extract specific datasets relevant to DURAARK.

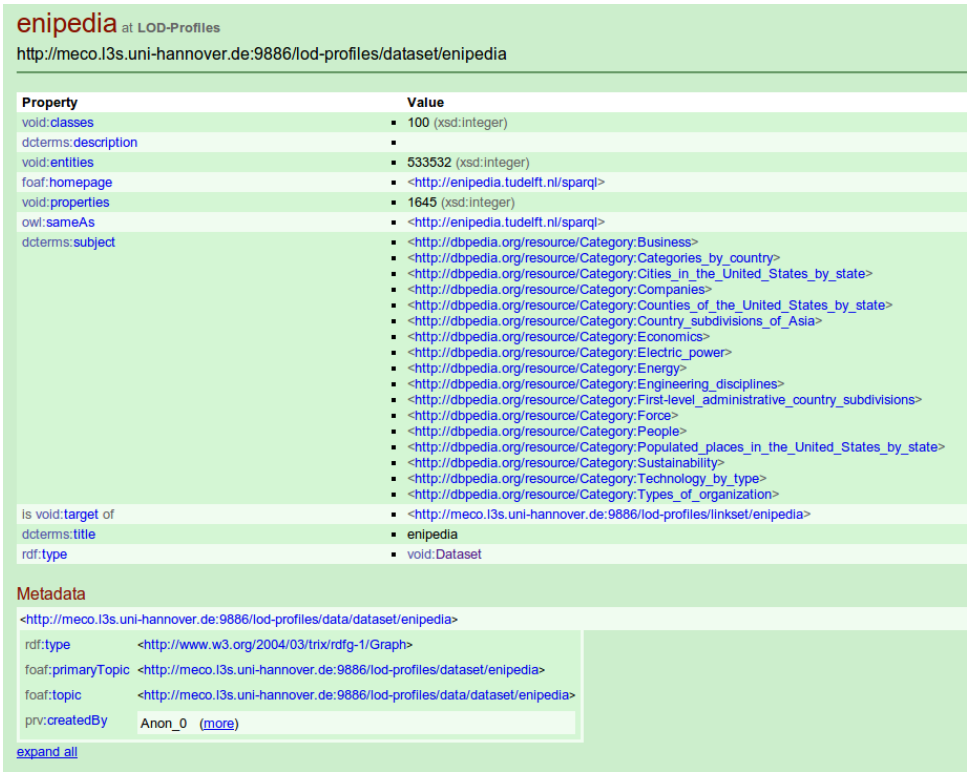

Figure 13: Dataset profile generated for the Enipedia dataset.

Figure 13 depicts an example dataset profile generated by our profiling platform. Enipedia[33] explores applications for wikis and the semantic web in the fields of energy and industry. The data is thereby of relevance to DURAARK. As can be seen in the figure, the description in the profile is quite useful to gauge the contents of the dataset and thus determine the relevance to a particular subject (in our case, DURAARK).

[33] http://enipedia.tudelft.nl/

5.2.1 Entity Recognition

The analysis of sampled resources for a set of datasets consists of an annotation process using Named Entity Recognition (NER) and disambiguation tools (DBpedia Spotlight[34]). From each resource we extract the textual content assigned to properties corresponding to the dataset. For instance, these can include {`dbprop:buildingType`, `rdfs:label`, `rdfs:comment`, `geo:geometry`, `foaf:isPrimaryTopicOf`, `skos:prefLabel`, `dcterms:description`, `dcterms:alternative`, `dcterms:title`, `owl:sameAs`, `dbprop:startDate`, `dbprop:completionData`, `foaf:name`, `dbpedia-owl:architect`}; and perform contextual, that is resource-wise, NER. This establishes a common descriptive layer of top-ranked entities for each dataset extracted from DBpedia.

As the NER process can pose a bottleneck, we introduce an *incremental annotation* extraction process to alleviate this issue. This process avoids annotating resources similar to previously annotated ones by reusing already obtained annotations. Thus, for a predefined threshold similarity τ, from a pool of existing annotations $\mathcal{A}$, we assign an annotation to a resource if the similarity (resource-annotation) computed by the Jaccard's index is above threshold τ:

$$\forall a \in \mathcal{A} : \; J(r, a) = \frac{|r \cap a|}{|r \cup a|} \tag{2}$$

where $a \in \mathcal{A}$ represents already extracted annotations, while r is a resource instance which is analysed using the *incremental annotation* process.

5.2.2 Category Annotation

From the extracted annotations (DBpedia entities) $\mathcal{A}$, we analyse the set of assigned categories for each annotation. Such information is extracted from the DBpedia graph via the property `dcterms:subject` representing the topic covered by an entity. Furthermore, we leverage the hierarchical category organisation (as defined by SKOS schema: `skos:broader` and `skos:related`) assigned to entities within DBpedia.

However, such information extracted about categories is only useful when ranked according to their relevance for each dataset. Hence, we compute a normalised *relevance score* for each category assigned to a dataset by taking into account (i). entities assigned to a category intra- and inter-datasets; and (ii). number of entities assigned to a dataset and

[34] `http://spotlight.dbpedia.org`

over all datasets, see Equation 3:

$$score(t) = \frac{\Phi(t, D)}{\Phi(\cdot, D)} + \frac{\Phi(t, \cdot)}{\Phi(\cdot, \cdot)}, \quad \forall t \in \mathcal{T} \wedge D \in \mathcal{D} \tag{3}$$

where $\Phi(\cdot, \cdot)$ represents the number of entities associated with a topic t and for a dataset D, in case of void arguments, it outputs the number of entities in a dataset or over all datasets.

5.2.3 Automated Annotation Validation & Filtering Approach

Validation and filtering of extracted annotations is necessary, due to noise inherited from NER&NED results. The approach we propose for filtering out noisy annotations takes into account the contextual support given for an annotation from the resource instance it is extracted from. Therefore, we compute a *confidence score* which measures the similarity between an annotation and a resource using Jaccard's index similar to Equation 2, based on values extracted from properties `dbpedia-owl:abstract` and `rdfs:comment`, and the set of analysed properties listed in Section 5.2.1, respectively.

Whereas, in the validation phase we consider only entities that have a *confidence score* above some pre-define threshold and use human evaluators to assess the relevance of an extracted annotation with respect to the resource context.

5.2.4 Results and Evaluation

Our current implementation includes datasets relevant to DURAARK. Our topic annotation used representative, randomly selected samples of resources from each datasets, with approximately 100 instances for each resource type. Steps included NER, category extraction and threshold-based filtering using our *relevance & confidence scores*.

From the extracted categories based on the resulting annotations, we incorporated only the top-*50* categories being the most representative ones for a dataset based on the computed *normalised-score*. Results obtained from this processing are stored as part of a VoID[35]-based dataset catalog currently being provided as part of our Data Observatory efforts[36].

The evaluation of annotation accuracy was measured based on two datasets: (a). annotation accuracy without any filtering (see Section 5.2.3); and (b). annotation accuracy after

[35] `http://www.w3.org/TR/void/`
[36] `http://data-observatory.org/lod-profiles`

filtering, where only annotations with scores above some threshold (in our case ≥ 0.15) are considered. The accuracy was measured for **1000** extracted annotations, picked randomly from $\mathcal{A}$. For (a) the accuracy was **71**%, whereas for (b) after filtering annotations below threshold $\tau \geq 0.15$. We observed an increase in accuracy of almost +**10**%.
Our demo application[37] focuses mainly on representation, profiling and search functionalities of the analysed datasets based on the structured descriptions. Figure 14 shows the interface of the dataset graph explorer.

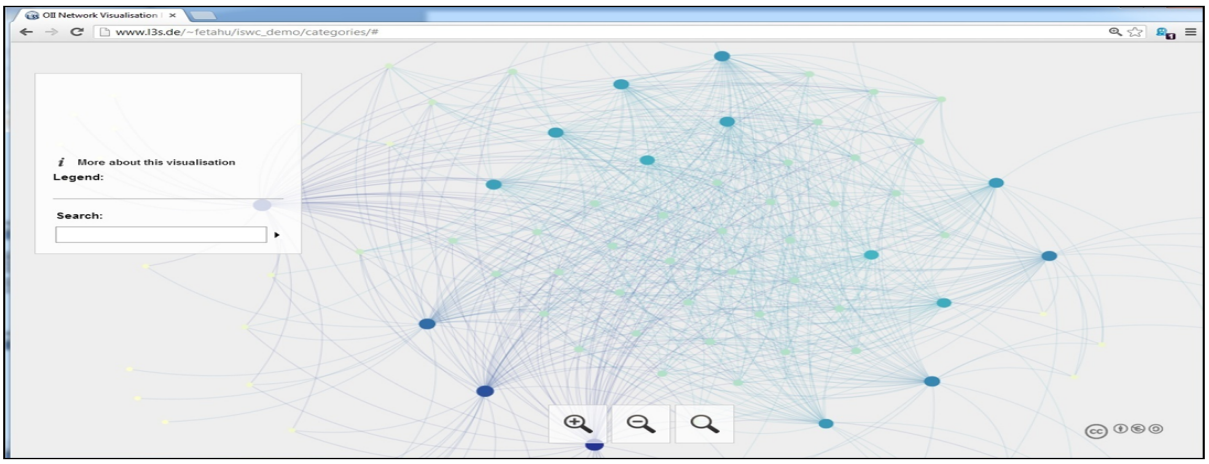

Figure 14: Screenshot of the graph explorer.

Figure 15 depicts the exploratory search functionality of datasets using extracted annotations and categories. The user interface provides the following:

- Exploratory search of datasets based on extracted annotations & categories
- Interlinking of datasets based on most representative categories
- List of ranked categories for each dataset

Our current processing pipeline is able to extract topic annotations for arbitrary Linked Data with only minimal manual intervention. Having applied it to a small subset of available datasets, our future work aims at the automatic profiling of all available LOD datasets, towards providing a more descriptive catalogue of Linked Datasets.

[37] http://data-observatory.org/lod-profiles/profile-explorer/

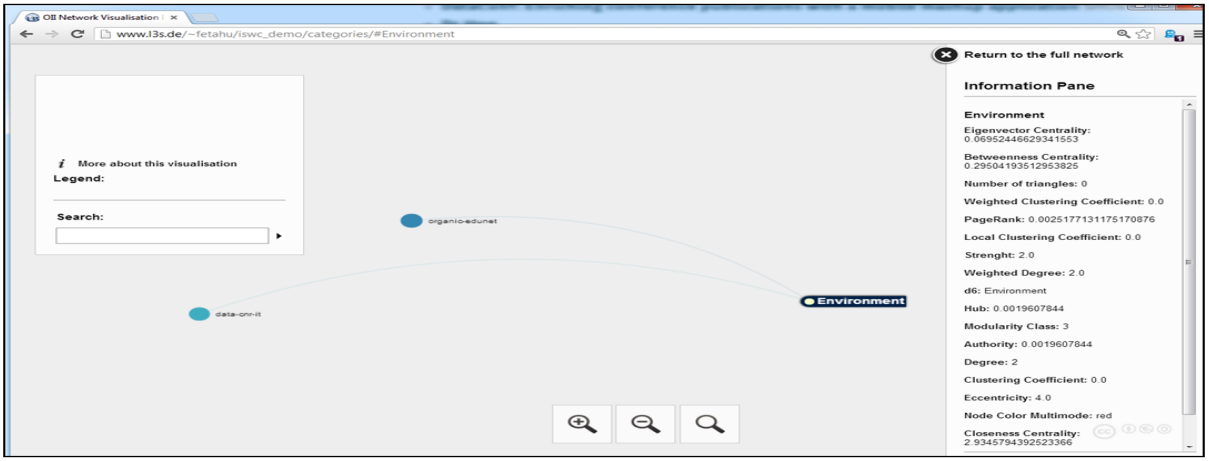

Figure 15: Screenshot of the graph explorer for example category "environment" interlinking different datasets shown on the right hand side panel.

6 Conclusions and future work

Based on work in previous stages of the DURAARK project and the requirements and use cases identified in work packages 2 and 7, in this report the general conceptual framework of the Semantic Digital Archive (SDA) and its subcomponent Semantic Digital Observatory (SDO) have been introduced. We have shown our approaches how BIMs that are semantically enriched with Linked Datasets can be stored in a OAIS-compliant Long Term Preservation (LTP) systems while also preserving the Linked Data itself. The main findings in this early stage of the project documented in this report can be summed up as follows:

- Using simple implementation agreements, legacy IFC models can be semantically enriched with arbitrary Linked Datasets without raising compatibility issues with existing commercial-of-the-shelf tools. Such techniques can be used for the controlled, formally rigid tagging of individual components in a BIM or metadata items for the description of Information Packages in OAIS-compliant preservation systems. (see section 3

- Potential datasets suitable for the enrichment of BIMs and archival metadata can be found, profiled, inventorized and mapped to other vocabularies using approaches shown in the Semantic Digital Observatory (SDO) of the DURAARK system. (see section 5)

- We have identified a number of conceptual approaches to archive such Linked Datasets alongside generic OAIS systems in a Semantic Digital Archive (SDA) to efficiently preserve evolving interlinked datasets that allow the reconstruction of arbitrary temporal states independent of the original online resources. (see section 4)

In subsequent activities of the DURAARK project a number of issues and research questions identfied in the individual sections of this report will be further elaborated and refined. The individual sub-processes and components described in this report will be implemented and software prototypes in the next months. These modules will be integrated into the overall DURAARK system as specified in WP 2.

References

[1] J. Beetz. Towards distributed, collaborative concept libraries for the construction industry. whitepaper circulated in the buildingSMART bSDD board and technical group. Technical report, Eindhoven University of Technology, 2013.

[2] J. Beetz. Towards a scalable network of concept libraries using distributed graph databases. In *Proceedings of CIB W78 2014*, Miami Floriday, Submitted for acceptance 2014.

[3] J. Beetz and B. de Vries. Building product catalogues on the semantic web. *Proc.CIB W78 Managing IT for Tomorrow*, page 221–226, 2009.

[4] J. Beetz, J. Van Leeuwen, and B. De Vries. IfcOWL: a case of transforming EXPRESS schemas into ontologies. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM*, 23(1):89–101, 2009.

[5] T. Berners-lee and D. Connolly. Delta: an ontology for the distribution of differences between rdf graphs. In *RDF Graphs. World Wide Web, http://www.w3.org/DesignIssues/Diff*, page 3, 2004.

[6] M. Bohms and F. Tolman. Building and construction eXtensible mark-up language (bcXML). In *IT in Construction in Africa 2001*, Mpumalanga, South Africa, June 2001.

[7] M. d'Aquin, A. Adamou, and S. Dietze. Assessing the educational linked data landscape. In *ACM Web Science 2013 (WebSci2013), Paris, France*. ACM, 2013.

[8] A. Ekholm. ISO 12006-2 and IFC–Prerequisites for coordination of standards for classification and interoperability. *Journal of Information Technology in Construction*, 10:275–289, 2005.

[9] B. Fetahu, S. Dietze, B. Pereira Nunes, D. Taibi, and M. Antonio Casanova. Generating structured profiles of linked data graphs. In *Proceedings of the 12th International Semantic Web Conference*. Springer, 2013.

[10] R. T. Fielding. *Architectural styles and the design of network-based software architectures*. PhD thesis, University of California, 2000.

[11] L. M. Giertz. *SFB and Its Development 1950-1980*. CIB/SFB International Bureau on behalf of the author; Foras Forbartha distributor, Dublin, Ireland, 1982. ISBN 0906120918.

[12] J. Hartmann, Y. Sure, P. Haase, R. Palma, and M. d. C. Suárez-Figueroa. OMV–ontology metadata vocabulary. In *ISWC 2005 Workshop on Ontology Patterns for the Semantic Web*, 2005.

[13] M. Hartung, J. Terwilliger, and E. Rahm. Recent advances in schema and ontology evolution. In Z. Bellahsene, A. Bonifati, and E. Rahm, editors, *Schema Matching and Mapping*, Data-Centric Systems and Applications, pages 149–190. Springer Berlin Heidelberg, Jan. 2011.

[14] ISO. 12006-3:2006 building construction — organization of information about construction works — part 3: Framework for object-oriented information, 2006.

[15] ISO 12006-2:2001 building construction – organization of information about construction works – part 2: Framework for classification of information. International standard, 2001.

[16] A. Khattak, K. Latif, S. Khan, and N. Ahmed. Managing change history in web ontologies. In *Fourth International Conference on Semantics, Knowledge and Grid, 2008. SKG '08*, pages 347–350, Dec. 2008.

[17] T. Liebich. IFC 2x edition 3 model implementation guide, 2009.

[18] C. Lima, J. Stephens, and M. Böhms. The bcXML: supporting eCommerce and knowledge management in the construction industry. *ITCon*, 8:293–308, 2003.

[19] C. Lima, A. Zarli, and G. Storer. Controlled vocabularies in the european construction sector: Evolution, current developments, and future trends. In *Complex Systems Concurrent Engineering*, pages 565–574. 2007.

[20] W. Miller and E. W. Myers. A file comparison program. *Software: Practice and Experience*, 15(11):1025–1040, 1985.

[21] J. Mynarz. Capturing temporal dimension of linked data. http://blog.mynarz.net/2013/07/capturing-temporal-dimension-of-linked.html.

[22] N. F. Noy, A. Chugh, W. Liu, and M. A. Musen. A framework for ontology evolution in collaborative environments. In *The Semantic Web-ISWC 2006*, page 544–558. Springer, 2006.

[23] N. F. Noy and M. Klein. Ontology evolution: Not the same as schema evolution. *Knowledge and information systems*, 6(4):428–440, 2004.

[24] D. Ognyanov and A. Kiryakov. Tracking changes in RDF(S) repositories. In A. Gómez-Pérez and V. Benjamins, editors, *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, volume 2473 of *Lecture Notes in Computer Science*, pages 373–378. Springer Berlin Heidelberg, 2002.

[25] R. Palma, P. Haase, O. Corcho, and A. Gómez-Pérez. Change representation for OWL 2 ontologies. 2009.

[26] D. W. G. E. Publishing. *DINI Certificate Document and Publication Services 2010*. DINI, version 3.0 edition, 2011.

[27] Technical specification SN/TS 3489. implementation of IFD library support in IFC., 2010.

[28] L. Stojanovic, A. Maedche, B. Motik, and N. Stojanovic. User-driven ontology evolution management. In A. Gómez-Pérez and V. R. Benjamins, editors, *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, number 2473 in Lecture Notes in Computer Science, pages 285–300. Springer Berlin Heidelberg, Jan. 2002.

[29] B. Swartout, R. Patil, K. Knight, and T. Russ. Toward distributed use of large-scale ontologies. In *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*, 1996.

[30] D. Taibi, B. Fetahu, and S. Dietze. Towards integration of web data into a coherent educational data graph. In *In Proceedings of 3rd LILE.*, 2013.

[31] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web and web 2.0 meet format risk management: P2 registry. In *iPres2009: The Sixth International Conference on Preservation of Digital Objects*, June 2009. Event Dates: October 5th and 6th, 2009.

[32] S. Törmä, J. Oraskari, and N. V. Hoang. Distributed transactional building information management. *LDAC 2012*, page 9, 2012.

[33] S. Tunnicliffe and I. Davis. Changeset, 2005.

[34] M. Vander Sande, P. Colpaert, R. Verborgh, S. Coppens, E. Mannens, and R. Van de Walle. R&Wbase: git for triples. In *Proceedings of the 6th Workshop on Linked Data on the Web*, May 2013.

A Overview and history of the buildingSMART Data Dictionary (bSDD)

The buildingSMART Data Dictionary (bSDD) started off as two independent initiatives in the Netherlands and in Norway. The earliest versions of LexiCon, the early version of the Dutch bSDD, started off as research activities at STABU in 1995. The Norwegian initiative was initiated as a result of one of the conclusions from a large research project named “Samspillet i byggeprosessen” (Collaboration in the building process) where lack of a common understanding of building terms and objects, was identified as one of the obstacles that needed to be overcome in order to improve the collaboration between the actors in a building process. A common ontology was believed to be the answer. At the time the oil industry was working on a similar problem in the EPISTLE project. Norway decided to do a full scale test by adding terms from wood and wooden products into EPISTLE by the help of POSC/CEASAR. The work started in 1998 and concluded in 1999 in the report “The wall testcase”. This was also the start of the "Bygg og Anlegg Referanse Bibliotek" (Building and construction reference library - BARBi) The conclusion was that EPISTLE would work but was far too complicated and work demanding. At that time Norway had also established contact with STABU and the LexiCon project. Together they initiated a group in ISO/TC59 with the purpose of creating a framework for data dictionaries as a subset of the EPISTE v.3.1 standard. The work started in 1999 and the name of the standard was ISO/PAS 12006-3 later nicknamed “IFD” (International framework for dictionaries). The work group behind ISO 12006-3 consist of representatives from ICIS, as well as experts from both LexiCon, BARBi as well as from IAI the creators of the IFC standard. It was already then obvious that IFD(bSDD) and IFC(bsDM) would be complementary standards that should get a tight integration in the future. This was at least obvious from the group developing the IFD(bSDD) standard. The first version of ISO 12006-3 (IFD/bSDD) was ready in 2002. Work on implementing the standard had already started and both BARBi and LexiCon claimed to be implementations of the ISO 12006-3 standard. While the work in Norway concentrated on creating tools and methodologies as well as an open API to access the data, the Dutch initiative was mostly about creating content. Norway (BARBi) had lots of content but mostly terms and their translations into English, German, Norwegian and French. This was imported from a database of terminology experts working on translating European standards into Norwegian. There was very few relationships in the BARBi library. But in 2003 Norway

started a project to define wooden products that also included establishing relationships between concepts.

The first public API to BARBi was published as a WSDL API (v.0.8) in 2005 followed by version 0.9 shortly after. At that time it had become clear that the benefits from bSDD/IFD would come from having one common library of terms with their GUIDs. It had at that point already started to appear several similar initiatives with different content not sharing anything else than being based on the same ISO 12003-6 standard. The standard itself only dealt with the structure of the content and did not really solve the major problem of harmonizing and comparing ontologies, dictionaries and classification systems from across domains and languages. The term "harmonize" is also quite difficult. Anyone that has attempted to harmonize two established classifications know that it's a near impossible task. Classification is like bit like religion in that sense. In an attempt to resolve this problem the BARBi project introduced the idea of a context. A context would allow people to create the structures they wanted inside IFD/bSDD and enforce any kind of ruleset to that context. That reduced the "harmonizing" problem to agreeing about the naming and rough classification of the concepts itself. A task that was believed to be much easier to agree about. After all two classification systems might be constructed by using the exact same words. It's how to organize the words in the structure that introduces most of the problems.

During the work of establishing BARBi , Norway made several attempts of importing the Dutch LexiCon data, but as LexiCon was a work in progress. Since there was no such thing as a GUID implemented in LexiCon that work was soon stopped as it introduced lots of duplications in the BARBi library. Instead the two countries decided to formalize cooperation and replace both LexiCon and BARBi with one unified IFD library. The work involved redesigning the API to fit the need of both parties as well as getting the content harmonized on concept level and a letter of intent was signed at 26 of January 2006. The official version of IFD library with its API in version 2.0 was officially released in September the same year in Lisbon. After that date several parties has joined the initiative and innumerous hours have been spent in trying to establish bSDD as an official part of buildingSMART and as "the" common structure for anyone working with ontologies for the building sector. From 2006 USA and Canada became partners representing buildingSMART USA, buildingSMART Canada. The list of observers also increased by many new members: ARCOM (USA), CCN (ZA), COBO systems (BE) Construction Information Ltd. (NZ), CRB (CH), GAEB/DIN (DE) NATSPEC (AUS), NBS (UK),

Norconsult (NO) and RTS (FIN).

Technical infrastructure On the technical side the API and technical infrastructure was gradually improved. API 2.0 was released in 2007 and API 3.0 in early 2011. These APIs where all SOAP based and while performing well they all had a problem when dealing with many simultaneous users. Early versions of the API also had lots of stability problems making it ill suited for any serious business need. As a long time partner in bSDD standardisation, development and implementation Catenda was heavily depending on bSDD for its products.

Content development Most of the original content of bSDD came from a database of technical terms used for translations of European standards into Norwegian. The original content consisted of roughly 9.000 concepts. This was mostly concepts without relationships to other concepts except from the standard where the term originated. But each concept had translations into French, German, English and Norwegian. The first version of LexiCon had roughly 1500 concepts and was imported in 2006 and in 2010 STABU made an attempt to import the latest version. Unfortunately this created a lot of duplicated content which has haunted bSDD since. On the positive side most LexiCon concepts had relations to other concepts. From 2009 and onwards Norway and Standard Norway added content related to building products. This content is considered rather complete and is also browsable through a logical structure. CSI has also contributed with content from the OmniClass standard and the whole of IFC's propertyset structure was imported in 2010. As of January 2014 bSDD consists of around 85.000 concepts with roughly 200.000 names. Mostly in Dutch, German, English (US, British, and Canadian as well as the common "International English) and Norwegian. There are now around 90.000 relationships in the database. Some in more than one context.

B Current bSDD and IFC enrichment compliant to SN/TS 3489

The current state-of-the-Art is based on the buildingSMART Data Dictionary bSDD which is discussed in section 3 and in more detail in appendix A. Reference and instantiation from within part 21 STEP Physical File (SPF)) IFC model instances can be carried out following the standard SN/TS 3489:2010 "Implementation of IFD Library support in IFC" [27]. In this document best practices are recommended to use the mechanisms provided in the IFC model to reference semantic definitions of objects in the IFD library implementation "bsDD" based on the ISO 12006:3 [14]. Since referencing approach proposed in the standard does not require the modification of the EXPRESS schema of the IFC model, they are (backwards) compatible with existing IFC tools and are readable using existing legacy tools. A number of applications have implemented dedicated support into their user interfaces to display information referencing external content according to this standard. The basic mechanism of SN/TS 2489 relies on the `IfcClassificationReference`, `IfcPropertyReferenceValue` and `IfcPropertyDependencyRelationship` entities in combinations with `IfcPropertySingleValue` instances which are collected and assigned to objects individually or via the typing mechanism. The key construction to allow the assignment of property/value pairs whose definition can and should be looked up in the external library is the use of to "magic token" keywords `IfdProperty` and `IfdValue` which are used in the `IfcPropertyReferenceValue` to signify the meaning of the `IfcClassificationReference` instances and how these should be interpreted.
To illustrate the mechanism standardized in SN/TS 3489, consider the following example snippet (listing 5), which is also illustrated further in figure 16. In this example and IfcPropertySingleValue with the name of "total height' and a value "10.0" with the (IFC) unit '.METRE.' is created and is assigned to an `IfcProxy` via an `IfcRelDefinesByProperties` relationship. In order to specify the actual meaning of the "total height" property a link is created to an entry in an external library that specifies its semantic meaning. The referenced concepts GUIDs refer to the concepts 'total height' (19JxIAb6qHtW00025QrE$V) and 'meter' (11lneAMgqHu000025QrE$V).

```
#8 = IFCSIUNIT(*, .LENGTHUNIT.,     , .METRE.);
#100 = IFCPROXY('3Fp4r0uuX5ywPYOUG2H2A4', #2, 'Proxy', 'Description of Proxy', $,
    #101, #51, .PRODUCT., 'Product proxy defined externally');
#300=IFCRELDEFINESBYPROPERTIES('35YdWMmwr4rQ61AZPsifP7',#2,$,$,(#100),#301);
```

```
#301=IFCPROPERTYSET('3Fp4r0uuX5ywPYOUG2H2A5',#2,
'Pset_With_IfcPropertyDependencyRelationship',$,(#310));

#310=IFCPROPERTYSINGLEVALUE('total height',$, IFCREAL(10.0),#8);

#311=IFCPROPERTYDEPENDENCYRELATIONSHIP(#310,#320,$,$,$);
#312=IFCPROPERTYDEPENDENCYRELATIONSHIP(#310,#321,$,$,$);

#320=IFCPROPERTYREFERENCEVALUE('IfdProperty',$,$, #330);
#321=IFCPROPERTYREFERENCEVALUE('IfdValue',$,$, #331);

#330=IFCCLASSIFICATIONREFERENCE('bsddpilot.catenda.no',
    '19JxIAb6qHtW00025QrE$V',$,#200);
#331=IFCCLASSIFICATIONREFERENCE('bsddpilot.catenda.no',
    '11lneAMgqHu000025QrE$V',$,#200);
```

Listing 5: Partial IFC file (in the SPF format) to demonstrate the semantic enrichment of engineering data following the SN/TS 3489 standard).

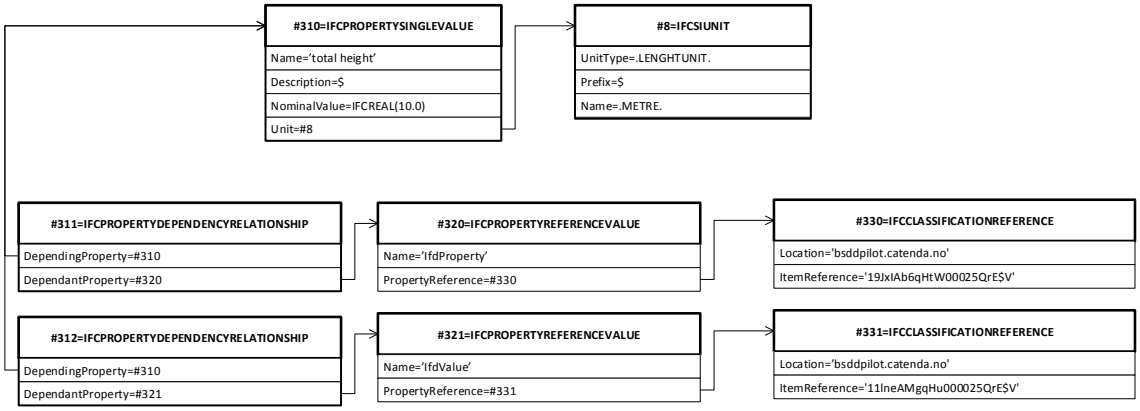

Figure 16: **Diagram illustrating the semantic enrichment suggested in SN/TS 3498 as provided by the example snippet in listing 5**

C Example fragment of the bSDD definition of a quay wall

```
<bsdd:3j9G00BS0Htm00025QrE$V>
    a       owl:Class ;
    rdfs:label "Quay wall of retaining wall elements"@en , "Kademuur van
        keerwandelementen"@nl-NL ;
    rdfs:subClassOf <http://www.tue.nl/ddss/duraark#3apCK0BS0Htm00025QrE$V> ,
        <http://www.tue.nl/ddss/duraark#3K5r60BROHtm00025QrE$V> ;
    ifd:conceptType "SUBJECT" ;
    ifd:guid "3j9G00BS0Htm00025QrE$V" ;
    ifd:status "DRAFT" ;
    ifd:versionDate "2010.03.12" ;
    ifd:versionId "1 2010.03.12" .
```

Listing 6: Fragment of the bSDD vocabulary showing the defition of the 'quay wall' concept serialized in the TURTLE format.

D Example data set of a versioned bSDD vocabulary using RDF named graphs

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:     <http://www.w3.org/2002/07/owl#> .



@prefix duraark_cs: <http://duraark.eu/vocab/changeset/schema#> .
@prefix bsdd : <http://buildingsmart.org/bsdd#> .
@prefix bsdd_version: <http://duraark.eu/ex/archive/bssd/versioning#> .
@prefix : <http://duraark.eu/ex/archive/bssd/versioning#> .


{
    :CS1  rdf:type duraark_cs:ChangeSet ;
      rdfs:label "CS1"^^xsd:string ;
      duraark_cs:createdDate
              "2013-12-01"^^xsd:date ;
      duraark_cs:creatorName
              "DURAARK Semantic Digital Observatory v 1.0"^^xsd:string ;
      duraark_cs:additions
              <http://duraark.eu/ex/archive/bsdd/changeset/1/additions> ;
      duraark_cs:removals <http://duraark.eu/ex/archive/bsdd/changeset/1/removals>
          .

    :CS2  rdf:type duraark_cs:ChangeSet ;
      rdfs:label "CS2"^^xsd:string ;
      duraark_cs:createdDate
              "2014-01-01"^^xsd:date ;
      duraark_cs:creatorName
              "DURAARK Semantic Digital Observatory v 1.0"^^xsd:string ;
      duraark_cs:precedingChangeSet
              :CS1 ;
      duraark_cs:additions
              <http://duraark.eu/ex/archive/bsdd/changeset/2/additions> ;

      duraark_cs:removals <http://duraark.eu/ex/archive/bsdd/changeset/2/removals>
          .

}

#initial dump of an external vocabulary
<http://duraark.eu/ex/archive/bsdd/initial> {
      <http://buildingsmart.org/bsdd#3vHRQ8oT0Hsm00051Mm008>
```

```
            a owl:Class ;

            rdfs:subClassOf <http://buildingsmart.org/bsdd#3vHdqCoT0Hsm00051Mm008> ,
                <http://buildingsmart.org/bsdd#3vHRQ8oT0Hsm00051Mm008> ;

            rdfs:comment "a complete unit consisting of a door frame and a door leaf
                or leaves,
            supplied with the essential hardware and wheatherseal, as a product from
                a single source"@en ,
             "A barrier to an entry that usually swings
            or slides to open and close the entry."@en ;

            rdfs:label "bloc-porte"@fr-FR ,  "Deur"@nl-NL ,
            "T\"{u}reinheit (von einem Anbieter)"@de-DE , "IfcDoor"@ifc-2X4 , "door
                set"@en , "door"@en ;


            bsdd:conceptType "SUBJECT" ;
            bsdd:guid "3vHRQ8oT0Hsm00051Mm008" ;
            bsdd:status "DRAFT" ;
            bsdd:versionDate "2012.11.05 13:57:25" ;
            bsdd:versionId "1" .
}


##### changes detected by a diff util in the SDO
<http://duraark.eu/ex/archive/bsdd/changeset/1/deletions>{
    <http://buildingsmart.org/bsdd#3vHRQ8oT0Hsm00051Mm008> rdfs:label
        "T\"ureinheit (von einem Anbieter)"@de-DE .
}

<http://duraark.eu/ex/archive/bsdd/changeset/1/additions>{
    <http://buildingsmart.org/bsdd#3vHRQ8oT0Hsm00051Mm008> rdfs:label
        "T\"ur"@de-DE .
}

###### changes detected in a subsequent run of the SDO diff tool
##### changes detected by a diff util in the SDO
<http://duraark.eu/ex/archive/bsdd/changeset/2/deletions>{
    <http://buildingsmart.org/bsdd#3vHRQ8oT0Hsm00051Mm008> rdfs:label
        "T\"ur"@de-DE .
}

<http://duraark.eu/ex/archive/bsdd/changeset/2/additions>{
    <http://buildingsmart.org/bsdd#3vHRQ8oT0Hsm00051Mm008> rdfs:label  "T\"ur mit
        Zarge"@de-DE .
}
```

Listing 7: Example query that demonstrates enriched building data that can be exploited.

Glossary

buildingSMART Data Dictionary bSDD The international reference repository of building related concepts governed by the buildingSMART organization. Based on the International Framework for Dictionaries (IFD).. 6, 13, 14, 16, 48, 53, 55

Building Information Model (BIM) Object-oriented, parametric and process-oriented data structures to organize information relevant to buildings. 5, 53

Industry Foundation Classes (IFC) The Industry Foundation Classes is an open interoperability model for the exchange of information reltated to building and construction. 5, 53

International Framework for Dictionaries (IFD) Conceptual framework and data model to organize building-related information. Formally specified in the ISO 12006 parts 2 and 3. Basis for the reference vocabulary. 6, 53

Linked Open Data (LOD) Set of accessible information published openly for reuse. The envisioned 'global graph of data' or 'cloud' employs Semantic Web technologies to cross-reference data across networks. Among its most important nuclei is the DBPedia data set that is derived from the WikiPedia corpus long. 53

LOD Linked Open Data long. 53

Long Term Preservation (LTP) The internationally widely accepted Open Archival Information System (OAIS) framework, defines preservation as

> "The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term."

(source: OAIS "Pink Book"). 6, 15, 16, 44, 53

Open Archival Information System (OAIS) Framework defining general concepts and best practises for Digital Long Term Preservation. 53

Resource Description Framework (RDF) Conceptual approach to modelling information. It is based on triples containing a 'subject' 'predicate' and 'object' that are described with Uniform Resource Identifiers (URI), most often reachable via web protocols such as HTTP. RDF models form directed graphs that can span across networked locations. Popular clear-text serialization formats include RDF/XML, N3 and Turtle. long. 53

Semantic Digital Archive (SDA) a part of the DURAARK framework that stores snapshots of linked data sets referenced from archives and their descriptions. 3, 7, 8, 44, 53

Semantic Digital Observatory (SDO) a part of the DURAARK system that crawls, fetches, monitores and updates external data sets stored for preservation in the SDA. 7, 8, 44, 53

STandard for the Exchange of Product data (STEP) A group of information standards covering a wide spectrum of engineering domains grouped under the ISO 10303 series of standards. 5, 6, 15, 53

STEP Physical File (SPF) A clear-text file format defined in the STandard for the Exchange of Product Data (STEP). Specified in the ISO 10303, part 21 and often referred to as "Part 21 file" or "SPFF" (STEP Physical File Format). 48, 53