Towards Preservation of semantically enriched Architectural Knowledge

Stefan Dietze[1], Jakob Beetz[2], Ujwal Gadiraju[1], Georgios Katsimpras[1],
Raoul Wessel[3], René Berndt[4]

[1] L3S Research Center, Leibniz University, Hannover, Germany
{dietze; gadiraju, katsimpras}@l3s.de
[2] Department of the Built Environment, Eindhoven University of Technology, The Netherlands
{j.beetz}@tue.nl
[3] Computer Graphics Group, University of Bonn, Germany
wesselr@cs.uni-bonn.de
[4] Fraunhofer Austria Research GmbH, Visual Computing, Graz, Austria
rene.berndt@vc.fraunhofer.at

Abstract. Preservation of architectural knowledge faces substantial challenges, most notably due the high level of data heterogeneity. On the one hand, low-level architectural models include 3D models and point cloud data up to richer building information models (BIM), often residing in isolated data stores with insufficient support for ensuring consistency and managing change. On the other hand, the Web contains vast amounts of information of potential relevance for stakeholders in the architectural field, such as urban planners, architects or building operators. This includes in particular Linked Data, offering structured data about, for instance, energy-efficiency policies, geodata or traffic and environmental information but also valuable knowledge which can be extracted from social media, for instance, about peoples' movements in and around buildings or their perception of certain structures. In this paper we provide an overview of our early work towards building a sustainable, semantic long-term archive in the architectural domain. In particular we highlight ongoing activities on semantic enrichment of low-level architectural models towards the curation of a semantic archive of architectural knowledge.

Keywords. Architecture, Semantic Web, Linked Data, Digital Preservation, Information Extraction, Building Information Model

1   Introduction

Long-term preservation of architectural knowledge - from 3D models to related Web data - faces a wide range of challenges in a number of use cases and scenarios, which are illustrated in Figure 1. In these diverse use-cases, preservation has to satisfy needs of a range of stakeholders, including architects, building operators, urban planners and archivists.

During the lifecycle of built structures, several engineering models are produced, updated and maintained, ranging from purely geometric 3D/CAD models and point clouds to higher level, semantically rich Building Information Models (BIM). Partial

domain models at different stages are highly interrelated and interdependent and include meronomic, spatial, temporal and taxonomic relationships. Apart from these BIM-internal explicit and implicit inter-relationships, a considerable number of references are also made to external information and data sets which imposes new challenges for digital long term preservation. For example, buildings to some degree can be considered as assemblies of various concrete building products which are specified by individual product manufacturers that have to be accessed in future maintenance, modification or liability scenarios.
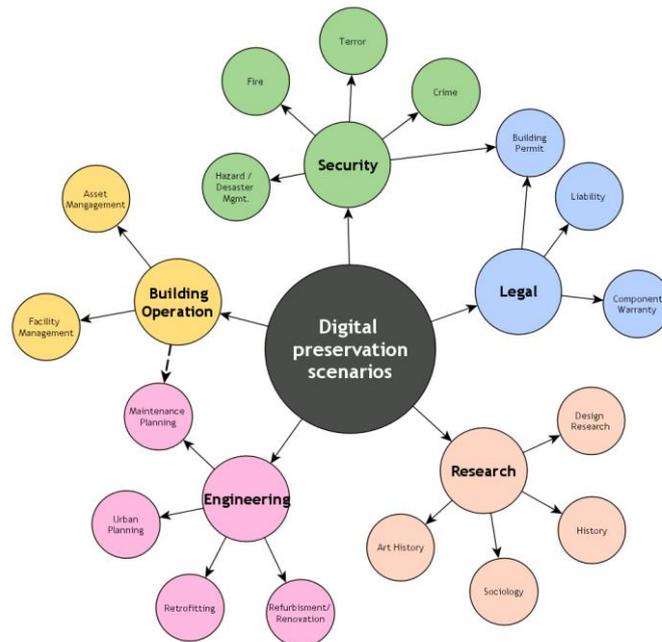
Fig. 1. Schematic overview of digital preservation scenarios & stakeholders

The individual building components and the building as a whole on the other hand have to comply with standards and local building regulations that are subject to constant evolvement and have to be preserved alongside the building model. Apart from such technical engineering information, the Web, in particular the Web of data and the social Web, contain an increasing amount of contextual information about buildings, their geo-location, history, legal context, the surrounding infrastructure or the usage and perception of structures by the general public. Examples include in particular the wide range of Linked Data [2] about geo-data[1], building-related policies[2] or traffic statistics[3] as well as the wide range of information which can be extracted from the frequency and content of social media, such as tweets or Flickr images, for in-

---

[1] For instance, http://www.geonames.org/ or https://geodacenter.asu.edu/datalist/

[2] For instance, energy efficiency guidelines at http://www.gbpn.org/databases-tools/building-energy-rating-policies

[3] A wide range of traffic and transport-related datasets at http://data.gov.uk

stance about the perception and use of buildings by the general public. Such information is distributed across the Web, is evolving constantly and is available in a variety of forms, structured as well as unstructured ones. Integration and interlinking as well as preservation strategies are of crucial importance. Particularly with regards to preservation, i.e. the long-term archival of all forms of architecturally relevant knowledge, challenges arise with respect to:

- *Semantic enrichment* of low-level architectural models
- *Interlinking & archiving* of related models (across different abstraction levels and model types, across different datasets and repositories including open data and manufacturer-specific data, covering evolution at different points in time, covering parts or related contexts of particular models)
- *Preservation & temporal analysis:* capturing and supporting the evolution of models, buildings and related data
- *Maintaining consistency* across archived data over time

The Web of (Linked) Data is a relatively recent effort derived from research on the Semantic Web, whose main objective is to generate a Web exposing and interlinking data previously enclosed within silos. The Web of Data is based upon simple principles based on the use of dereferencable HTTP URIs, representation and query standards like RDF, OWL [1] and SPARQL[4] and the extensive use of links across datasets. While Linked Data (LD) principles have emerged as de-facto standard for sharing data on the Web, our work is fundamentally aiming at (a) creating a semantic digital archive (SDA) for the architectural domain according to LD principles and (b) leveraging on the existing wealth of Web data, particularly Linked Data, to gradually enrich the archive. Given the distributed evolution of all considered knowledge and data types, dedicated archiving and preservation strategies are of crucial importance.

In this paper, we introduce our current vision and future work within the recently started project DURAARK ("Durable Architectural Knowledge")[5], aimed at the long-term preservation of low-level architectural models gradually enriched with higher level semantics. The archived models are described as a part of a well-interlinked knowledge graph which in particular incorporates the temporal evolution of building structures and their contexts. We introduce an early draft of the overall architecture together with the semantic enrichment components. One of the requirements for preservation of structured Web data is dataset curation – i.e. profiling and classification of available datasets into their coverage (geographical, topics, knowledge types). We introduce our research activities on curation, aiming at generating catalogs (and archives) of available datasets useful to the architecture and construction sector and other interested parties.

---

[4] http://www.w3.org/TR/rdf-sparql-query/
[5] http://www.duraark.eu

2   Durable Architectural Knowledge - Approach & Overview

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 information standards* as well as the *enrichment and correlation of architectural models with related Web data*. This approach applies to both, the building models as well as interlinked data. While earlier efforts where focused on the preservation of proprietary, binary file formats such as Autodesk's DWG and DXF on a byte stream level [8][10], DURAARK makes distinct use of open, text-based formats from the family of ISO 10303 standards , referred to as STEP – Standard for the Exchange of Product data. In particular, the Industry Foundation Classes (IFC) [9] model along with its open specifications published and governed by the buildingSMART organization[6], 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 additional set of several hundred standardized properties that can be attached to individual entity instances and can conveniently be extended, providing a meta-modeling facility to end-users and software vendors alike.

Most, commonly IFC models are serialized as Part 21 – SPFF (STEP Physical File Formats) [14] and to a lesser extent as a content-equivalent XML representation following ISO 10303 part 28 [15]. These formats (albeit using different model schemas) have also been chosen in long-term preservation scenarios in other engineering domains [7][11], and have earlier been identified as most promising candidates for future research endeavors [8][10]. The self-documenting clear-text encoding of both instance files and schemas increase the likelihood of future reconstruction and make them less error-prone on physical levels of bit-rotting. Next to the aforementioned part 21 and 28 serializations, the DURAARK project will also provide an RDF representation of these models[12][13], which allows easier semantic enrichment and integration with other archival process chains. The architecture of the DURAARK system can be roughly divided into three layers:

1. *Processing tools* that help users to semantically and geometrically enrich and prepare architectural models for ingestion.
2. A *Semantic Digital Archive* that provides a common registry, access and preservation facility for enriched BIM and related Web data.
3. An *OAIS[7] compliant archival system* that maintains AIPs consisting of IFC files, RDF graphs of the linked data used for semantic enrichment as well as compressed and uncompressed point cloud data sets to document as-build states of documented buildings. This will be implemented on top of the existing state-of-the-art archival products such as Rosetta.

---

[6] http://buildingsmart.org
[7] http://public.ccsds.org/publications/archive/650x0m2.pdf

One of the key features of the DURAARK approach is the gradual enrichment of low-level architectural models. Enrichment starts with *geometric enrichment*, which produces structured metadata (IFC, BIM) out of low-level architectural models and scans. Based on such structured metadata, *semantic enrichment* aims at retrieving higher-level semantic information about the described structure, for instance, about its geolocation, history or surrounding infrastructures. All data, low-level models as well as the enriched metadata will be archived in an OAIS-compliant preservation system.
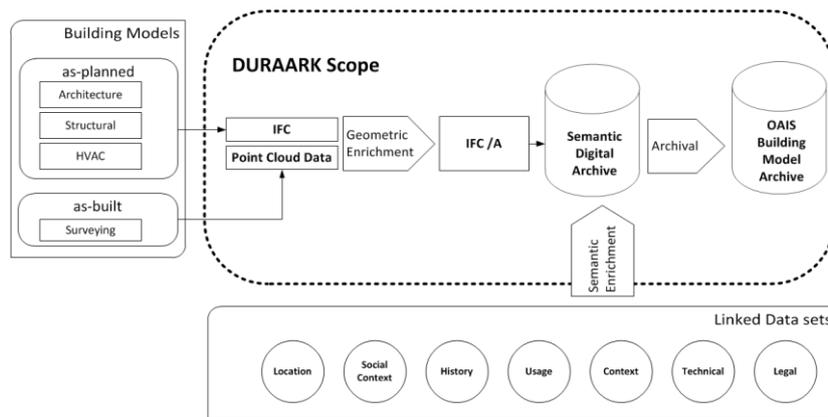
Fig. 2. DURAARK architecture overview

3 Semantic Enrichment & Preservation of Architectural Knowledge

As part of the preprocessing tools to be developed for the ingestion and preservation of building information models, an essential component facilitates the *semantic enrichment* (see [23]) and annotation as well as the extraction and compilation of relevant metadata. Semantic enrichment exploits both, *expert-curated domain models* and *heterogeneous Web data sources*, in particular Linked Data, for gradually enriching a BIM/IFC model with related information. The metadata enrichment aims to populate BIM according to the schema shown in Figure 3 and includes:

1. *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 adaption due to their limited exposure via standard interfaces. This includes, for instance, the buildingSMART Data Dictionary (bsDD) exposing several tens of thousands of concepts. While currently limited to custom SOAP and REST web services, the DURAARK project will expose this information as 5 star Linked Data preserved as part of the SDA.

2. ***Automated & manual interlinking and correlation with related Web data:*** as part of this step, architectural models (IFC/BIM) will be enriched with related information prevalent on the Web, for instance, about the geolocation (and its history), surrounding traffic, transport and infrastructure and the usage and perception by the general public. Building on previous work on entity linking [4], data consolidation and correlation for digital archives [6][5], dedicated algorithms for the architectural domain will be developed, for instance tailored to detect data relating to specific geospatial areas or to specifically architecturally relevant resource types. Additionally, *during the ingestion for archival* which will be 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 graph-based yet 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 has to be tackled in an integrated fashion.

Generally, while within the LD graph, in theory all datasets (and RDF statements) are connected in a way, 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 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[8]) while on the other hand, other datasets are meant to change highly frequently (for instance, Twitter feeds or *Highways Agency Live Traffic Data*[9]). 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[10]

---

[8] http://www.data.gov.uk/dataset/2008_injury_road_traffic_collisions_in_northern_ireland

[9] http://www.data.gov.uk/dataset/live-traffic-information-from-the-highways-agency-road-network

[10] http://semantic.eurobau.com/

or DBpedia[11]. Depending on the specific requirements, nature and dynamics of individual datasets, we are exploring *Web data preservation strategies*, including (a) non-recurring capture of URI references to external entities as is common practice within the LD community, (b) non-recurring archival of subgraphs or the entire graph of the external dataset, (c) periodic crawling and archiving of external datasets.
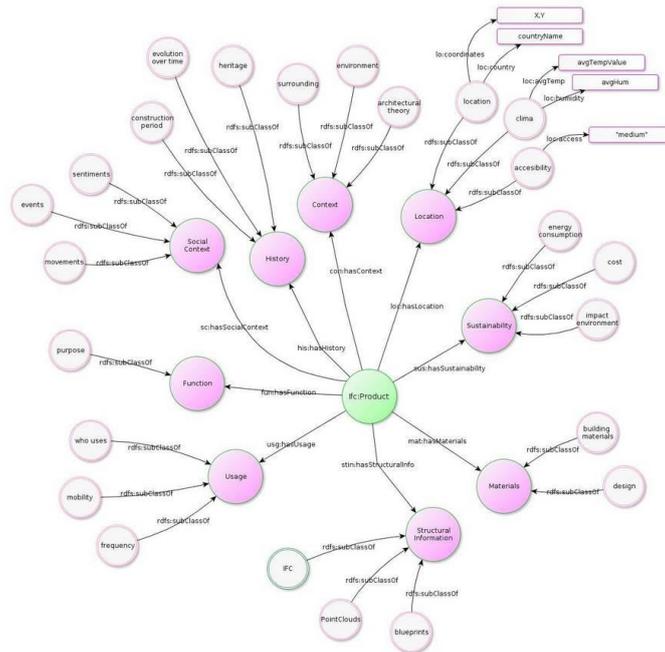
Fig. 3. Enrichment data schema

In order to facilitate informed decisions about suitable preservation strategies for individual datasets, additional structured information about the characteristics of each dataset is required, what is addressed through dedicated *data curation strategies*.

4   Curation and Preservation of Datasets and Vocabularies

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, which 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 instance, whether the data contains building-related policy information or traffic or environmental data). For this purpose we are currently in the process of establish-

---

[11] http://dbpedia.org

ing 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

4.1 Towards semi-automated generation of a dataset observatory & archive for the architectural domain

While there exists a wealth of relevant Web datasets, particularly Linked Data, providing useful data of relevance to the architectural field (see Figure 4 for examples), metadata about available datasets is very sparse.
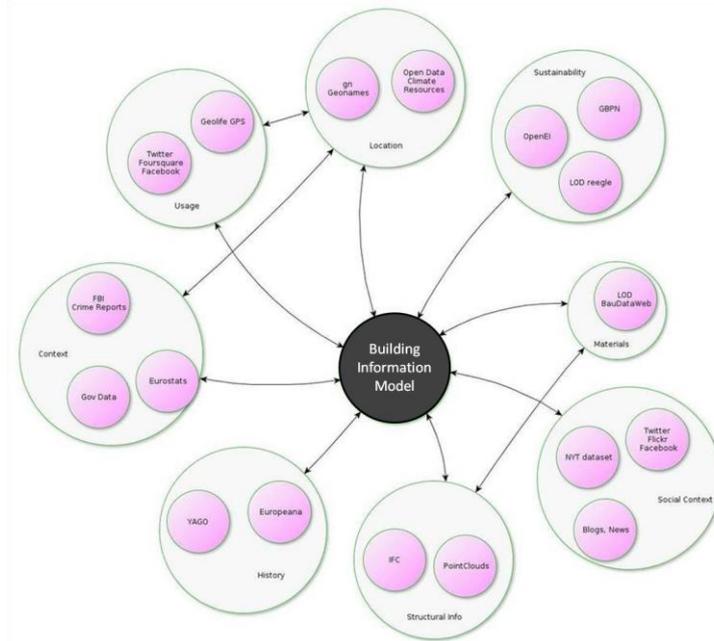
Fig. 4. Overview of building related knowledge types and datasets

Considering LD and Open Data in general, the main registry of available datasets is the DataHub[12], currently containing over 6000 open datasets and, as part of the Linked Open Data group[13], 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,

---

[12] http://datahub.io
[13] http://datahub.io/group/lodcloud

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 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 achivists 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"[14]) has been set up (though not yet populated) to collect datasets of relevance to the architectural field. While the DataHub is based on CKAN[15], 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, by
  - 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* (see [22][23])

As part of such activities, we are currently in the process of generating a *structured dataset catalog, which adopts VoID[16]* 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 [3] and follows similar aims as the work described in [24], yet we aim to not only provide metadata about the dynamics of datasets but also additional metadata ab,out for instance, 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, for instance, whether it describes traffic statistics for the Greater London area or energy efficieny policies for Germany.

---

[14] http://datahub.io/group/linked-building-data (recently founded group on the DataHub)
[15] http://ckan.org/
[16] http://vocab.deri.ie/void

As described in Section 3, different preservation strategies are considered for each dataset, depending on the dynamics and frequency and size of updates. While each strategy requires knowledge about the datasets to interact with, for instance, the URI of their SPARQL endpoints, our VoID-based "Linked Building Data" catalog will provide the basis for realising such individual preservation strategies and will be enriched with preservation-related metadata, for instance about the update procedures and evolution of each dataset.

4.2 Expert-based curation of domain vocabularies

In the past, a number of research efforts have aimed at providing manually curated, structured vocabularies of the various building-related engineering domains. Among them are the EU-projects eConstruct [16], IntelliGrid [18] and SWOP [17], as well as other national and international initiatives such as FUNSIEC [19]. The buildingSMART data dictionary (bsDD)[17] has the ambition to be a central vocabulary repository that allows the parallel and integrated storage of different vocabularies such as the various classification systems (OMNICLASS Masterformat[18], UNICLASS[20], or SfB(-NL[19]) which are widely adopted in the respective countries to structure building data. The bsDD also servers as the central repository to store meta-model extensions of IFCs - referred to as PSets - which are not part of the core model schema but are recognized as typical properties of common building component. A number of commercial domain-specific building product catalogs and conceptual structures have been established that are captured in propriatary data structures that are not yet exposed as Open Data, yet have gained the status of de facto industry standards. These include the international ETIM[20] classification for the description of electronic equipment in buildings, the Dutch Bouwconnect[21] platform, the German Heinze[22] product database and the CROW library for infrastructural objects[23]. Such structured vocabularies are often tightly integrated and oriented at local builing regulation requirements and best practices and are often underlying structures for ordering higher-level data sets such as standardized texts for tendering documents (the German StLB[24], the Dutch STABU system[25], Finnish Haahtela[26] etc.)

Even though their use and application in the context of the Semantic Web and LD has been suggested time and again [21], the uptake of harmonized structures is still in its infancy although internationally anticipated by large end-user communities.

---

[17] http://www.buildingsmart.org/standards/ifd
[18] http://www.csinet.org/Home-Page-Category/Formats/MasterFormat.aspx
[19] http://nl-sfb.bk.tudelft.nl
[20] http://e5.working.etim-international.com
[21] http://www.bouwconnect.nl
[22] http://www.heinze.de/
[23] http://www.gww-ob.nl/
[24] http://www.stlb-bau-online.de/
[25] http://www.stabu.org
[26] https://www.haahtela.fi/en/

5 Discussion and future works

In this paper we have presented an overview of the current and future work within the DURAARK project for creating a semantic digital archive for the building and architecture domain. While the project is in its early stages, currently focusing on gathering requirements and designing initial prototypes for the main components, our main contributions are the proposed architecture for digital preservation of architectural knowledge, the semantic enrichment approach and our currently ongoing work towards curation of architecturally relevant Web datasets, which builds the foundation for implementing tailored, specific and efficient strategies for preservation of continuously evolving Web datasets.

Our future work will be dedicated to fully realising our data curation approach by creating a structured dataset catalog containing meaningful metadata of architectural-related datasets. This will form the basis to implement (a) enrichment and interlinking algorithms which gradually enrich Building Information Models and (b) to fully realise preservation strategies which will enable to assess and analyse the temporal evolution of architectural models as well as correlated Web data.

Acknowledgments

References

[1] Antoniou, G., van Harmelen, F. Web Ontology Language: OWL.in S. Staab, & R. Studer (eds.) Handbook on Ontologies, pp. 67-92, 2004,

[2] Bizer, C., T. Heath, Berners-Lee, T. (2009). Linked data - The Story So Far. Special Issue on Linked data, International Journal on Semantic Web and Information Systems.

[3] D'Aquin, M., Adamou, A., Dietze, S., Assessing the Educational Linked Data Landscape, ACM Web Science 2013 (WebSci2013), Paris, France, May 2013.

[4] Nunes, B. P., Dietze, S., Casanova, M.A., Kawase, R., Fetahu, B., Nejdl, W., Combining a co-occurrence-based and a semantic measure for entity linking, ESWC 2013 – 10th Extended Semantic Web Conference, Montpellier, France, May (2013).

[5] Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y., Senellart, P., Exploiting the Social and Semantic Web for guided Web Archiving, The International Conference on Theory and Practice of Digital Libraries 2012 (TPDL2012), Cyprus, September 2012.

[6] Dietze, S., Maynard, D., Demidova, E., Risse, T., Peters, W., Doka, K., Stavrakas, Y., Entity Extraction and Consolidation for Social Web Content Preservation, in Proceedings of 2nd International Workshop on Semantic Digital Archives (SDA), Pafos, Cyprus, September 2012.

[7] prEN 9300-003:2005, 2005. LOng Term Archiving and Retrieval of digital technical product documentation such as 3D, CAD and PDM data. PART 003: Fundamentals and Concepts.

[8] Smith, M., 2009. Curating Architectural 3D CAD Models. International Journal of Digital Curation, 4(1), pp.98–106.

[9]   ISO 16739:2013 Industry Foundation Classes, Release 2x, Platform Specification (IFC2x Platform).

[10] Berndt, R. et al., 2010. The PROBADO Project - Approach and Lessons Learned in Building a Digital Library System for Heterogeneous Non-textual Documents. In M. Lalmas et al., eds. Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 376–383.

[11] VDA, 2006. VDA 4958 Long term archiving (LTA) of digital product data which are not based on technical drawings.

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

[13] Pauwels, P. et al., 2011. Three-dimensional information exchange over the semantic web for the domain of architecture, engineering, and construction. AI EDAM, 25(Special Issue 04), pp.317–332.

[14] ISO 10303-21:2002 Industrial automation systems and integration -- Product data representation and exchange -- Part 21: Implementation methods: Clear text encoding of the exchange structure.

[15] ISO 10303-28:2007 Industrial automation systems and integration -- Product data representation and exchange -- Part 28: Implementation methods: XML representations of EXPRESS schemas and data, using XML schemas,

[16] Tolman, F. et al., 2001. eConstruct: expectations, solutions and results. Electronic Journal Of Information Technology In Construction (ITcon), 6, pp.175–197.

[17] Böhms, M. et al., 2009. Semantic product modelling and configuration: challenges and opportunities. , 14, pp.507–525.

[18] Dolenc, M. et al., 2007. The InteliGrid platform for virtual organisations interoperability. , 12, pp.459–477.

[19] Lima, C. et al., 2006. A framework to support interoperability among semantic resources. In Interoperability of Enterprise Software and Applications. Springer, pp. 87–98

[20] Crawford, M., 1997. UNICLASS: Unified Classification for the Construction Industry, RIBA Publications.

[21] Beetz, J. & de Vries, B., 2009. Building product catalogues on the semantic web. Proc.CIB W78 "Managing IT for Tomorrow", pp.221–226.

[22] Paes Leme, L. A. P., Lopes, G. R., Nunes, B. P., Casanova, M.A., Dietze, S., Identifying candidate datasets for data interlinking, in Proceedings of the 13th International Conference on Web Engineering, (2013)

[23] Taibi, D., Fetahu, B., Dietze, S., Towards Integration of Web Data into a coherent Educational Data Graph, in Leslie Car, Alberto H. F. Laender, Bernadette F. Lóscio, Irwin King, Marcus Fontoura, Denny Vrandeèiæ, Lora Aroyo, José Palazzo M. de Oliveira, Fernanda Lima, Erik Wilde (editors), Companion Publication of the IW3C2 WWW 2013 Conference, May 13–17, 2013, Rio de Janeiro, Brazil. IW3C2 2013, ISBN 978-1-4503-2038-2

[24] Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A., Observing Linked Data Dynamics, in the Proceedings of the 10th Extended Semantic Web Conference (ESWC2013), Montpellier, France, 26–30 May, 2013.