D2.2.2 System Architecture & Specification v1.0

DURAARK

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

Date: 2013-07-31
Version 1.0
Document id. : duraark/2013/D.2.2.2/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/2013/D.2.2.2
Title of document	:	D2.2.2 System Architecture & Specification v1.0
Deliverable type	:	Report
Contractual date of delivery	:	2013-07-31
Actual date of delivery	:	2013-07-31
Lead beneficiary	:	FhA
Author(s)	:	Thomas Bähr <thomas.baehr@tib.uni-hannover.de> (LUH) Jakob Beetz <J.Beetz@tue.nl> (TUE) René Berndt <rene.berndt@vc.fraunhofer.at> (FhA), Stefan Dietze <dietze@l3s.de> (LUH) Dag Fjeld Edvardsen <dag.fjeld.edvardsen@catenda.no> (Catenda) Ujwal Gadiraju <gadiraju@l3s.de> (LUH) Michelle Lindlar <michelle.lindlar@tib.uni-hannover.de> (LUH) Sebastian Ochmann <ochmann@cs.uni-bonn.de> (UBO), Martin Tamke <martin.tamke@kadk.dk> (CITA) Richard Vock <vock@cs.uni-bonn.de> (UBO)

Responsible editor(s)	:	René Berndt <rene.berndt@vc.fraunhofer.at> (FhA), Eva Eggeling <eva.eggeling@vc.fraunhofer.at> (FhA)
Quality assessor(s)	:	Stefan Dietze <dietze@l3s.de> (LUH), and Michelle Lindlar <michelle.lindlar@tib.uni-hannover.de> (LUH) Raoul Wessel <wesselr@cs.uni-bonn.de> (UBO)
Approval of this deliverable	:	
Distribution	:	Public
Keywords list	:	Quality Assurance, Risk Management Plan

Executive Summary

This deliverable presents the first version of the system architecture of DURAARK system. It describes the philosophy, decisions, constraints, justifications, significant elements, and any other overarching aspects of the system that shape the design and implementation. It will be complemented by deliverable D2.2.3 System Architecture & Specification v2.0 in month 12.

Table of Contents

Glossary

AIP Archival Information Package.

API Application programming interface.

BIM Building Information Modeling.

CAD Computer-aided design.

DIP Dissemination Information Package.

DPS Digital Preservation System.

GUI Graphical user interface.

HTTP Hypertext transfer protocol.

HVAC Heating, Ventilation, and Air Conditioning.

IDM Information Delivery Manual.

IFC Industry Foundation Classes.

JSON JavaScript Object Notation.

OAIS Open Archival Information System.

OWL Web Ontology Language.

PUID PRONOM Persistent Unique Identifier.

RDF Resource Description Framework.

REST Representational state transfer.

SDA Semantic Digital Archive.

SIP Submission Information Package.

SOA Service-oriented architecture.

SPARQL SPARQL Protocol And RDF Query Language.

STEP Standard for the exchange of product model data.

URI Uniform resource identifier.

XML eXtensible Markup Language.

1 Architectural goals and philosophy

The DURAARK system architecture is based on a component based approach. An individual software component is a software package, a web service, a web resource, or a module that encapsulates a set of related functions (or data). All system processes are placed into separate components so that all of the data and functions inside each component are semantically related (just as with the contents of classes). Because of this principle, it is often said that components are modular and cohesive. The component based approach allows the reuse of the developed components for the various use-cases described in D2.2.1. They can easily be rearranged to adopt to new usage scenarios, which will be identified during the projects duration.

Figure 1 illustrates an overview of the individual parts and workflow of the project that lead to the ingest of information. The three main activities can be distinguished:

- **Semantic enrichment of data** which allows the addition of information from various information sources to the building information models themselves as well as to the meta-data of the archival information package (AIP).
- **Geometric enrichment of data** which allows the inclusion of surveyed data in the form of point-clouds as additional as-built information alongside existing explicit building information models or as the main form of geometric representation where such explicit models are not available at the time of archival.
- **Preservation of data** which ensures that the information is available and useable for the designated community. It addresses the aspects of bit preservation, logical preservation and semantic preservation, which were described in detail in D2.2.1.

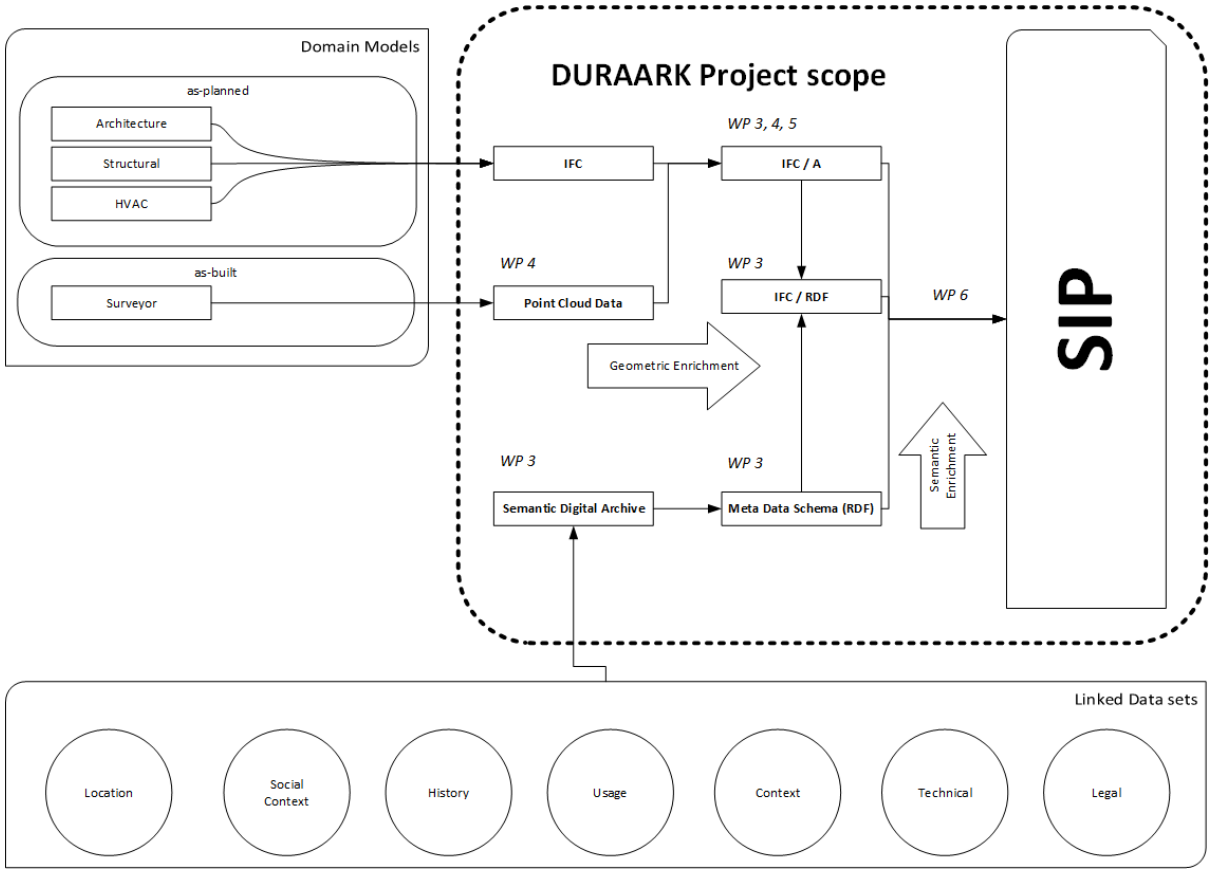

Figure 1: Systems overview of the DURAARK ingest.

2 Architecturally significant requirements

The functional and non-functional requirements have been described in deliverable D2.2.1.

3 Decisions, constraints, and justifications

3.1 Quality and organization of measured data

The methods developed in the course of the DURAARK project target usage scenarios from the Architecture, Engineering, Construction and Facility Management (AEC/FM) domain where a certain degree of precision and quality of computational results is mandatory. This imposes certain requirements on the input data and therefore the assumption is made that the measured point-cloud data is of a rather high quality (i.e. scans from high-quality laser scanning devices instead of consumer hardware). In general, each scan has to be (almost) spherical (that is, rays are cast from a point into all directions) and that the points are organized into rows and columns. Note that this is not a hard assumption as professional laser scanning devices capture the data organized in a fixed-size 2D-array.

Furthermore, in order to yield a satisfactory construction of a BIM model from point-cloud data, all relevant parts of the building have actually be "seen" by the scanner. Where small gaps in the data and scan shadows are a natural part of 3D scanning, good results cannot be guaranteed if large parts of the building are missing in the data. The same applies for parts of buildings that are not scanned at all.

For the format of point-cloud data, the E57 standard[1] has been selected. The main reasons for choosing this file format are that it is vendor-neutral, versatile and well-documented, as well as the availability of an open-source library (libE57[3]) for parsing it. The format can not only store the point data but also the scanner locations as well as images taken in the course of the scanning process. Scans from multiple locations may be contained in a single file.

The availability of the sensor locations can be a valuable hint for later processing and analysis of the point-cloud data. The images taken from the scanner's location are important for the computer vision methods that will be employed in WP5.

3.2 Quality and organization of BIM data

It is assumed that the data captured in a Building Information model is delivered with information on at least geometrical level that complies to the IFC model specification[1]. In particular files should be encoded as ISO 10303 part 21 files (referred to as SPF-STEP Physical File Format, which is also used in other engineering domains and their digital preservation systems[25]. Since the initial formal requirements of IFC models are very lax (about 80% percent of all schema-level attributes are optional) additional constraints will have to be fulfilled to qualify an IFC identified for archival. Such constraints are expressed in so-called Model View Definitions (MVD). An MVD for the minimal requirements on IFC files is being defined in the context of the DURAARK project. This MVD referred to as "IFC/A" also imposes constraints on the meta-data extracted from the IFC model such as authorships, software versions involved in the production of the model, measures and units and other information that will be exposed for indexing and searching in the AIPs. Flawed, corrupted or damaged files are not part of data that is considered suitable for ingest to a long term archive.

The IFC format will be the way to deliver Building Information Modeling data to the archive. The format has become the industry standard in the building industry for exchange of building and construction data. As it is a neutral and open specification it is not controlled by a single vendor or group of vendors. The DURAARK consortium has good contacts to buildingSMART group[2] (formerly the International Alliance for Interoperability, IAI) that maintains the data model of IFC and is continuously developing the format in order to facilitate interoperability in the architecture, engineering and construction (AEC) industry. The IFC model specification is open and available. It is registered by ISO and is an official International Standard ISO 16739:2013[3]

Together with the CAD software market and the general technological progress in the Construction industry the format is under constant development. Hence the projected interoperability between software packages using the same IFC file is not always given. These incompatibility problems cannot be part of the considerations of the DURAARK project.

[1] http://buildingsmart-tech.org/specifications/ifc-overview/
[2] http://www.buildingsmart.org/bim
[3] http://www.buildingsmart.org/openbim

3.3 Use of Semantic Web and Linked Data standards

The Semantic Web[9] is fundamentally based on the idea of machine-readable Web data (complementing unstructured Web documents) which add a layer of semantics to be exploited by search engines and, most notably, reasoning engines. Efforts on realizing the Semantic Web have led to a wealth of ontology and markup languages, most notably OWL[5] which exploit formal knowledge representation and description logics to conceptualize formal ontologies which allow inference through dedicated reasoning engines. 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 four simple principles, known as the Linked Data principles, which essentially dictate that every piece of data should be given a de-referenceable HTTP URI which, when looked up, should offer useful information using standards like RDF and SPARQL[11]. Additionally, data should be linked to other relevant resources therefore allowing humans and computers to discover additional information and to ensure a high level of interoperability through the use of shared vocabularies. The Web of Data, initially mostly an academic endeavor, is gradually capturing the attention of companies and institutions some of which have already taken steps towards making use of these technologies. Examples of this latest trend are for instance, the acquisition of Metaweb by Google, the use of Semantic Web technologies in public services provided by the BBC , and the release of governmental data following Linked Data principles by countries like the United Kingdom [4] or the United States [5]. Linked Data techniques have been applied to improve interoperability in a range of scenarios, such as service and API integration[17], educational resources interoperability[16], general data integration and linking[23], and interoperability in preservation and archiving contexts[15]. To this end, DURAARK assumes a high suitability of Linked Data-based approaches and technologies for sharing and interlinking of architectural knowledge and meta-data, as generated for instance within WP3.

Past and ongoing activities to harness these technologies also play an important role in the building and construction industry, including the architectural domain. 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

[4] http://data.gov.uk
[5] http://data.gov

EU-projects eConstruct[26], IntelliGrid[18] and SWOP[10], as well as other national and international initiatives such as FUNSIEC[21]. The buildingSMART data dictionary (bsDD) 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[6], UNICLASS[13], or SfB(-NL)[7]) which are widely adopted in the respective countries to structure building data. The bsDD also serves 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 proprietary 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[8] classification along with its commercial implementation in the 2BA platform[9] for the description of electronic equipment in buildings, the Dutch Bouwconnect[10] platform, the German Heinze[11] product database and the CROW library for infrastructural objects[12]. Such structured vocabularies are often tightly integrated and oriented at local building regulation requirements and best practices and are often underlying structures for ordering higher-level data sets such as standardized text for tendering documents (Geraman StLB[13], Dtuch STABU[14], Finnish Haahtela[15] etc.) Even though their use and application in the context of the Semantic Web and Linked Open Data has been suggested time and again[6], the uptake of harmonized structures is still in its infancy although internationally anticipated by large end-user communities.

The recent evolution and wide-spread adoption indicates that Linked Data is currently establishing itself as the de-facto standard for data integration on a Web scale. Although certainly in harmony with Semantic Web research, work on Linked Data has focused on applying lighter technologies, e.g., RDF(S) and SPARQL, on devising simple, yet

[6]http://www.csinet.org/Home-Page-Category/Formats/MasterFormat.aspx
[7]http://nl-sfb.bk.tudelft.nl/
[8]http://e5.working.etim-international.com/
[9]http://www.2ba.nl/
[10]http://www.bouwconnect.nl/
[11]http://www.heinze.de/
[12]http://www.gww-ob.nl/
[13]http://www.stlb-bau-online.de/
[14]http://www.stabu.org/
[15]https://www.haahtela.fi/en/

extensible and integrated vocabularies, and on establishing the fundamental means for exposing and interlinking data. Effectively these technologies have simplified the integration of heterogeneous data sources to a certain extent, providing common languages for data representation and querying, as well as by borrowing Web standards for uniquely and globally identifying entities and transporting data.These very aspects make these technologies particularly appealing for capturing, sharing and integrating architectural knowledge within DURAARK, which is likely to be highly heterogeneous, provided by diverse actors, and yet need to be effectively integrated on a global basis. Therefore, DURAARK meta-data will exploit Linked Data techniques and datasets to ensure a high level of interoperability and availability of generated data and meta-data. In particular, DURAARK will deploy existing state-of-the-art Linked Data storage and processing environments, which will be adopted and expanded to the needs of the architectural domain.

3.4 Digital preservation system: Rosetta

The OAIS compliant digital preservation system integrated into the DURAARK system architecture is Rosetta by Ex Libris[16]. Rosetta has been in productive use at TIB (LUH) since January 2012. TIB uses the system for their own data and furthermore hosts the consortially operated system for the two other German national subject libraries - the German National Library of Medicine (ZB MED) and the German National Library of Economics (ZBW). Together, the three libraries form the Goportis consortia, which conducted a digital preservation pilot study from 2009 through 2011. Within the pilot study several digital preservation systems were analyzed and a pilot implementation of Rosetta was tested against pre-defined criteria in a cooperatively operated digital preservation system. The Goportis consortia placed a high focus on the openness and the extendability of the system. Rosetta is a format agnostic system and is easily extendable for new workflows and materials through available APIs as well as its innate rule-based workflow engine and ability to support plugins. As a trustworthy, OAIS compliant digital preservation system it does, however, have a number of fundamental principles which hold true for all workflows and cannot be changed. As such, Rosetta implements the PREMIS data

[16]http://www.exlibrisgroup.com/category/RosettaOverview

model[17] to capture the information flow as well as information about intellectual entities, rights, agents, objects and events.

Within this context, an object is the smallest discrete unit of information. One or more objects make up an intellectual entity (IE), which is an intellectual unit for management and description of data in the digital preservation system. Furthermore, within the IE one or more objects may be grouped together as representations and objects may be further differentiated as files or bitstreams. The DURAARK system architecture needs to treat objects inline with the PREMIS data structure. The components should support output to which can be mapped to the PREMIS entities - especially agents and events.

Inline with the OAIS definition, a submission information package (SIP) can contain one or more intellectual entities. It should be decided within the system architecture whether both cases - n:1 and 1:1 intellectual entities per SIP should be realized. Within Rosetta complex (n:1) SIPs will be transformed into separate AIPs. Submission information can enter the Rosetta system in two ways: they are generated as part of a standard Rosetta deposit UI or they are created by ingest tools which function as a submission application. Such a submission application must follow a predefined Rosetta SIP structure. The submission application calls the Rosetta deposit through an API provided by Ex Libris.

Each AIP in Rosetta contains one intellectual entity and is stored in a METS XML file[20], which references the objects of the intellectual entity stored in the file system. meta-data to be generated by the DURAARK system components and to be included in the AIP must fit into the METS XML schema. It is assumed by LUH (TIB) that no DURAARK components will externally trigger an update process on a SIP, DIP or AIP kept within Rosetta. Preservation actions such as migration or a re-run of format identification fall within the governance of LUH (TIB) and will only be performed within the repository.

Rosetta will act as a "Light archive", meaning the data is available to users in addition to being part of the content repository. The other option - a "dark archive" where access to the data is highly or completely restricted - has been dismissed, because of the need of an additional repository layer.

Since Rosetta does not contains sophisticated user interfaces and content-based retrieval techniques for architectural 3D data, the decision was made to use PROBADO3D[8]

[17] http://www.loc.gov/standards/premis/

as a front-end for accessing the data preserved within Rosetta. PROBADO3D is one reference domain of the PROBADO project, which is a research effort to develop and operate advanced Digital Library support for non-textual documents[7]. PROBADO3D will be used as the interface for browsing and searching the archived data within Rosetta. This can either be done by using the PROBADO3D web pages or the various SOAP requests provided by the PROBADO3D system.

4 Layers or architectural framework

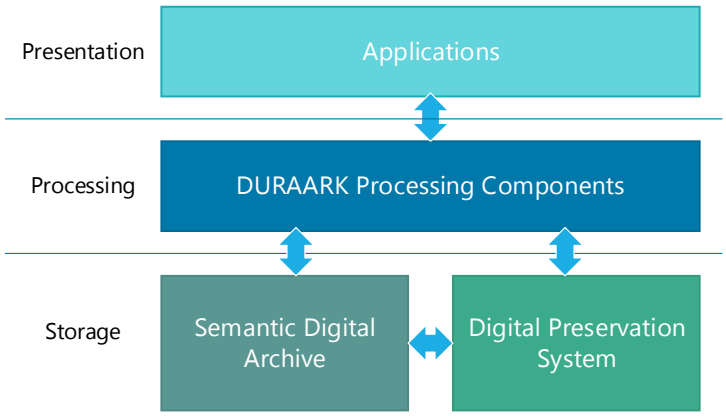

Figure 2: Overview of the logical packages of the DURAARK system.

This section gives an overview of the logical structure of the various components, which will be developed within the DURAARK project. Following the three-tier architecture pattern, the components are structured into three logical layers (see figure 2):

- **Presentation** This layer represents the components responsible for displaying information and results from the processing layer.
- **Processing** Components related to the geometric and semantic processing of the digital 3D architectural data.
- **Storage** This layer consists of two subsystems - `Semantic Digital Archive` and `Digital Preservation System` -, which are responsible for population and curation of building-related external vocabularies and for storage, retrieval and preservation planning.

The system architecture of DURAARK is based on the service-oriented architecture design paradigm (SOA). SOA is the aggregation of components that satisfy a business need and communicate between each other by exchanging structured messages following standardized APIs. It comprises components, services, and processes. Components are binaries that have a defined interface (usually only one), and a service is a grouping of components (executable programs) to get the job done. This higher level of application development provides a strategic advantage, facilitating more focus on the business

requirement[19]. The communication between the various components will be based on REST-based interfaces - where feasible - exposing either JSON or XML or both as message formats. Components involved in the exchange of large quantities of data (e.g. ingest of a point-cloud data) require more low-level and closer integration. The interfaces/protocols needed for these components will be part of D2.2.3.

Figure 3 shows the DURAARK stack, which illustrates the hierarchy of the various layers.

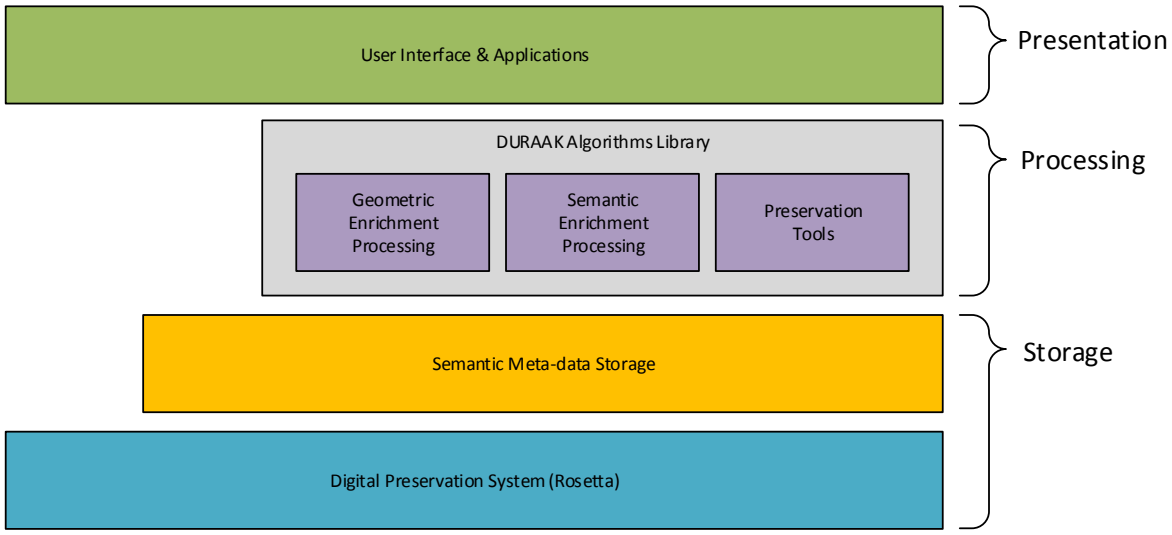

Figure 3: DURAARK Stack

4.1 User interface & applications

User interfaces and applications will allow users to (a) archive and modify data and (b) to query and consume data from the archive. While (a) is particularly addressing users like archivists and data providers, (b) is particularly addressing users such as architects or urban planners, who want to satisfy information needs as specified in the deliverable D2.2.1. It is intended that the semantic meta-data stored in the semantic digital archive provides higher-level search and query mechanisms which allow to retrieve data and models relevant to specific search criteria (e.g. for specific geo-locations, time periods, structures affected by certain energy-efficiency policies, structures within particular areas modified in certain periods etc).

During the ingest of new (point-cloud) data, the data needs to be pre-processed. This includes the generation of IFC from the point-cloud data and compression. The IFC generation in particular is a semi-automatic process in which the user may guide the algorithms with manual input where necessary or desired. For this purpose, a user interface will be developed to give the user a preview of the results of the extraction algorithm and to give her or him the ability to tweak the results. This interface lies between the input of the unprocessed data (either in the form of files or in the form of data read from a database, e.g. PROBADO3D) and the storage of the processed data (e.g. Rosetta).

A user interface for accessing and viewing the data stored in the archival system will be developed. This interface may include the difference-detection algorithms of WP4. For instance, a user may load two different datasets (two point-clouds or a point-cloud and a BIM model) and use difference finding and highlighting tools on them.

4.2 DURAARK algorithms library

4.2.1 Geometry enrichment processing

The algorithms developed in WP4 and WP5 deal with the processing and analysis of point-cloud data as well as BIM models. They include algorithms for the registration (alignment) of two representations of the same building, the detection and extraction of differences between them as well as compression methods. Also, algorithms for the generation of IFC data from given point-clouds and the transfer of semantics from a given IFC model to a point-cloud will be included.

These components (see section 5) will be compiled as libraries which may then be used in interactive or automated tools. The advantage of this model is that the functions implemented in the libraries are reusable and not necessarily tied to a specific program.

4.2.2 Semantic enrichment processing

The IFC model is a semantically rich data model that contains a range of information items that are interesting for storage in the meta-data set of an AIP that allow targeted searches and retrievals. Among them are information items on actors, geo-locations, tools and time stamps on individual objects, space allocations etc. Dedicated and customizable

extraction algorithms and tools are developed in the DURAARK project that allow to compile a rich set of meta-data from the BIM/IFC models. While some of this work will be based on own developments of the project partners themselves[22], other tools that e.g. measure model metrics of the ingested files will be based on work in similar directions contributed by other authors.

4.2.3 Preservation tools

In building and construction projects, a wide range of tools is used across a large number of engineering disciplines. This results in a great variety and heterogeneity of the data sets produced in a BIM as a final, merged results of all sub-models stemming from the various stakeholders. In addition to these cross-disciplinary interoperability issues, the different tools applied within the boundaries of a single domain (e.g. architecture) along with their associated data modeling approaches, geometry types and storage formats increase this heterogeneity of BIMs further. Even though interoperability models and data standards such as the IFCs - which are widely recognized as the most established open standard - address and partly overcome this dilemma by harmonizing and embracing a broad range of the most commonly used component descriptions, many issues still remain unsolved. One of the most pressing issues in reaching seamless interoperability is the lack of strictness of the underlying model schema. While this allows more flexibility and lowers the bar for mappings from proprietary in-application data structures, it also makes the subsequent processing of these models more challenging. For example, about 80 % of the ca. 2000 entity attributes are optional according to the formal schema. From a preservation point of view this means, that essential information about individual objects or the whole building model might not be present at ingest time (and can hence not be indexed for searches) while still being syntactically and semantically valid IFC model instances. To overcome this, two main approaches can be identified to further unify the resulting models:

- **Informal implementation agreements** limit the degrees of freedom of data modeling further. The collectively edited document which is officially recommended by the buildingSMART organization serves as a best practices guideline that is also taken into consideration for the (semi-automated) certification of individual implementations.

- **Model View Definitions (MVD)** are subsets of the extensive global model that allow the specification of e.g. 'required', 'obligatory' and 'disallowed' properties for individual exchange scenarios in a formally rigid way. The most well-defined and widely-used MVDs include the 'Coordination View' and 'Facility handover' views. While such views are standardized and public accessible through the standardization body, they envisaged future use also extends the application to individual per-project scenarios.

To allow reliable and trustworthy long term preservation of ingested models, the DURAARK project will develop such an MVD for archival purposes, referred to as "IFC/A" in a similar fashion as the PDF/A sub-standard. This IFC/A MVD determines the minimal information required e.g. by the meta-data extraction preprocessing steps described in 4.2.2 in accordance to the meta-data schema developed in WP3. In addition of the validation of the ingested models in accordance to the formally defined IFC/A MVD (which is described in an existing, standardized XML schema referred to as mvdXML), validation of ingested files also comprises checks on the self-containment of lined data referenced from within the BIM/IFC model. This latter step while guarantee that each external reference can be resolved using local archived copies of the respective data sets (e.g. bsDD references) at future retrieval points.

4.3 Semantic meta-data storage

DURAARK will develop and populate a large knowledge base of semantically enriched meta-data, which will complement BIM data and provides more expressive meta-data about the built structures, their context, environment, usage and history. Following the design decisions for exploiting Linked Data techniques and standards at the meta-data level, DURAARK meta-data storage will be based on graph-based RDF storage systems, providing SPARQL-based access for communication. While it is assumed that (a) large quantities of data have to be managed and (b) frequent queries need to be supported from other components, DURAARK will rely on established storage and SPARQL endpoint solutions, which will selected according to the requirements elicited in early stages of the project. SPARQL endpoints are services that enable users to query an RDF/OWL knowledge base via SPARQL. RDF triplestores, i.e. databases providing persistent storage and access to RDF graphs, usually provide SPARQL endpoints. In this sense, some of the

most widely used triplestores (e.g. Virtuoso, AllegroGraph, Joseki, Sesame/OpenRDF or Mulgara just to name a few), implement SPARQL endpoints (often compliant with different SPARQL versions/releases).

As Cheung[12] commented on his work, SPARQL helps solving interoperability problems derived from the underlying different technologies of each triplestore. Thus, SPARQL allows datasets in each triplestore to be accessed via standard SPARQL queries issued by clients to a common SPARQL endpoint service, an approach which allows creating cross-links at programming level[12]. From a linked data point of view, the recent availability of front-end tools for SPARQL endpoints such as Pubby is remarkable, as they allow creating linked data interfaces to SPARQL endpoints. In any case, an increasing number of triplestores (e.g. Virtuoso) provide native linked data exposure as part of their functionality. The importance and widespread use of SPARQL can be shown that DBpedia – according to many one of the more famous parts of the Linked Data efforts – provides a public SPARQL endpoint which enables users to query the RDF data-source with SPARQL queries.

4.4 Digital preservation system (Rosetta)

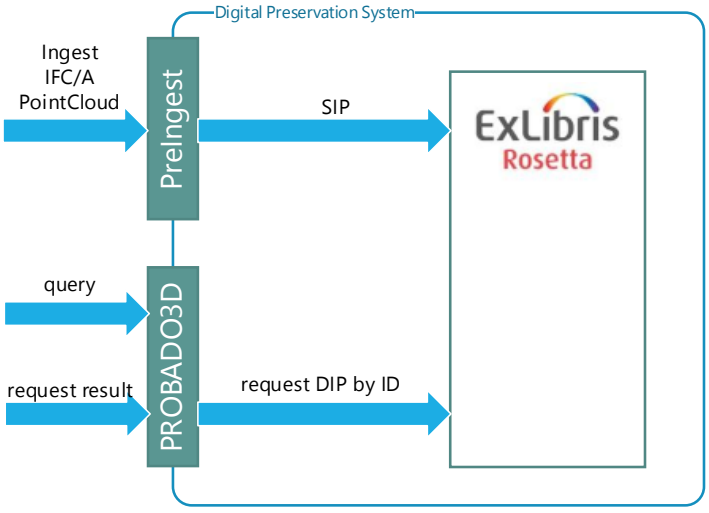

Figure 4: Overview of the Digital Preservation subsystem.

Rosetta is a scalable, format agnostic digital preservation system developed by Ex Libris in collaboration with the National Library of New Zealand and an international peer review group. It is built on open standards of the digital preservation community, such as METS[20] and PREMIS[4], and is extendable through available submission and dissemination APIs and SDKs. The system can be further enhanced through plug-ins for standard digital preservation tasks such as format characterization (file format identification, technical meta-data extraction, format validation), migration or risk extraction. This enables the digital preservation operator to easily deploy standard community tools like jhove[2] as well as internally developed tools for specific file formats, such as IFC. Figure 4 shows the components of the Digital Preservation subsystem. The PreIngest component accepts IFC/A and point-Cloud data from the DURAARK client tools and creates the SIP (Submission Information Package) for the OAIS-compliant Rosetta system. Rosetta will function as a light archive. PROBADO3D will function as the discovery layer - access to the object will be facilitated through an URL in PROBADO3D which points to the appropriate representation within the DIP in Rosetta. Within the DURAARK project the necessary changes for connection DURAARK and PROBADO3D will be addressed. Since the IFC file format is not supported by PROBADO3D, the system will be extended accordingly.

5 Components

This section gives an overview of all components to be developed within the elements of the DURAARK stack(see figure 3).

5.1 User interface & applications

Difference analysis visualization component
This component shall provide an interactive visualization for the comparison result of the difference analysis component (see below). This will be realized either as a standalone application or as a plugin to a third party software (e.g. Autodesk Revit).

- **INPUT** Annotated point-clouds or a custom file format for storing differences (similar to textual "diff" tools).
- **OUTPUT** None.

5.2 DURAARK algorithms library

5.2.1 Geometric enrichment processing

Registration component
This component shall provide functions for registering (aligning) multiple point-cloud scans and IFC files to each other. It would be most meaningful to integrate this into an interactive GUI environment, offering the user an interactive preview and controls over the registration process.

- **INPUT** A pair of two (possibly roughly pre-aligned) representations of a building (point-cloud or IFC model).
- **OUTPUT** A transformation aligning both representations.

Difference analysis component
This component shall provide functions for comparing two point-clouds or a point-cloud

with a given IFC model or legacy CAD data or between two IFC models. This library could be used in an interactive GUI environment that allows the user to see differences between the models with appropriate highlighting of parts that differ in the available representations. It should also be possible to extract and store differences in an appropriate file format (for instance an enriched, annotated point-cloud storing the correspondences and differences and a report highlighting the minimal, maximal, and median deviations).

- **INPUT** Either two point-clouds or a point-cloud and an IFC model/legacy CAD data.
- **OUTPUT** Annotated point-clouds or a custom file format for storing differences (similar to textual "diff" tools).

Compression component

This component shall provide functions for compressing point-cloud data. This library could, for example, be used in a "headless" executable that batch-processes input data or it could be used in an interactive application, offering the user a preview of the compressed data.

- **INPUT** Point-clouds and – if available – a corresponding IFC model serving as additional information for compression.
- **OUTPUT** Either point-cloud data that has been decimated in a meaningful way or data in a custom compressed format.

IFC extraction component

This component shall provide functions for creating IFC files from point-clouds in a (semi-)automatic manner. It may be used in an interactive GUI-environment, providing a multi-phase conversion workflow for creating an IFC model from given scans. The user should be able to intervene in each phase of the extraction in order to tweak or correct it where necessary while being supported by automatic algorithms where possible.

The scope of the extraction will be the coarse building structure consisting of floors, walls, doors and windows. This representation also enables to search for entities that can

be described in a simple manner (for instance, to find all instances of a door of a given width and height).

- **INPUT** A registered point-cloud consisting of multiple scans (including the information about the scanner positions in the global coordinate system).
- **OUTPUT** An IFC model containing the extracted entities (floors, walls, doors, etc.).

IFC-based geometric point-cloud augmentation component

The process of comparing a scan with an existing IFC opens up the possibility to connect semantically meaningful parts of the point-cloud with their corresponding IFC entities.

One possible approach might be: A decision which point belongs to which entity is made based on the distance between the point and a geometric representation of the entity. Semi-automatic feature detection will help identifying objects, e.g. walls that are positioned further off than a given threshold value. Some degree of manual intervention will be needed using a GUI.

- **INPUT** IFC model and matching point-clouds.
- **OUTPUT** Either an annotated point-cloud where each point is associated with its IFC entity (if any) or an enriched IFC model where each entity is assigned its corresponding subset of points.

Hidden structure detection component

This component will identify power sockets and outlets, as well as light switches and (if visible) distribution sockets and cavity sockets. The result will be a set of images with markups for the identified components (e.g. a light switch) and their probability. These markups are the input for shape grammars. The grammar represents the rules, according to which power and water lines are installed; the markups in the images are the terminal symbols. A 3D structure will be generated, which represents the known inputs (e.g. light switches, respectively terminal symbols) best. This component may be used in an interactive GUI-environment.

- **INPUT** Multiple separate point-clouds from different scanner positions (including the information about the scanner positions in the global coordinate system) and the images created during the scanning process.
- **OUTPUT** An IFC file containing the detected structures (power lines, etc.).

5.2.2 Semantic enrichment processing

Meta-data extraction from IFC model component
A considerable number of information facets that is relevant to indexing and searching of information in an archival context is already available in the IFC model itself. In accordance to D3.3.1 this component will parse the relevant information in the prepared IFC model and map them into meta-data sets that are exposed to the archival system. This meta-data includes authorship, software tools involved in the creation of the data, building components and their manufacturers and the intended functions of spaces

- **INPUT** IFC model.
- **OUTPUT** Meta-data schema instance enriched with information extracted from the IFC model.

Semantic enrichment (context) of BIM component
This component allows the access to external data sets of architectural relevance (geodata, energy-efficiency policy, related transport, related infrastructure etc) and the creation of links between entities in these data sets and individual building elements, spaces or the building as a whole. This component focuses on non-engineering data that is publicly available, extracted from unstructured Web information or added by e.g. librarians or archive curators such as the classification of architectural styles, social and historic context information and public perceptions(see figure 5).

- **INPUT** IFC model.
- **OUTPUT** Meta-data schema instance enriched with information extracted from the IFC model and from additional external resources stored / mirrored in the Semantic Digital Archive.

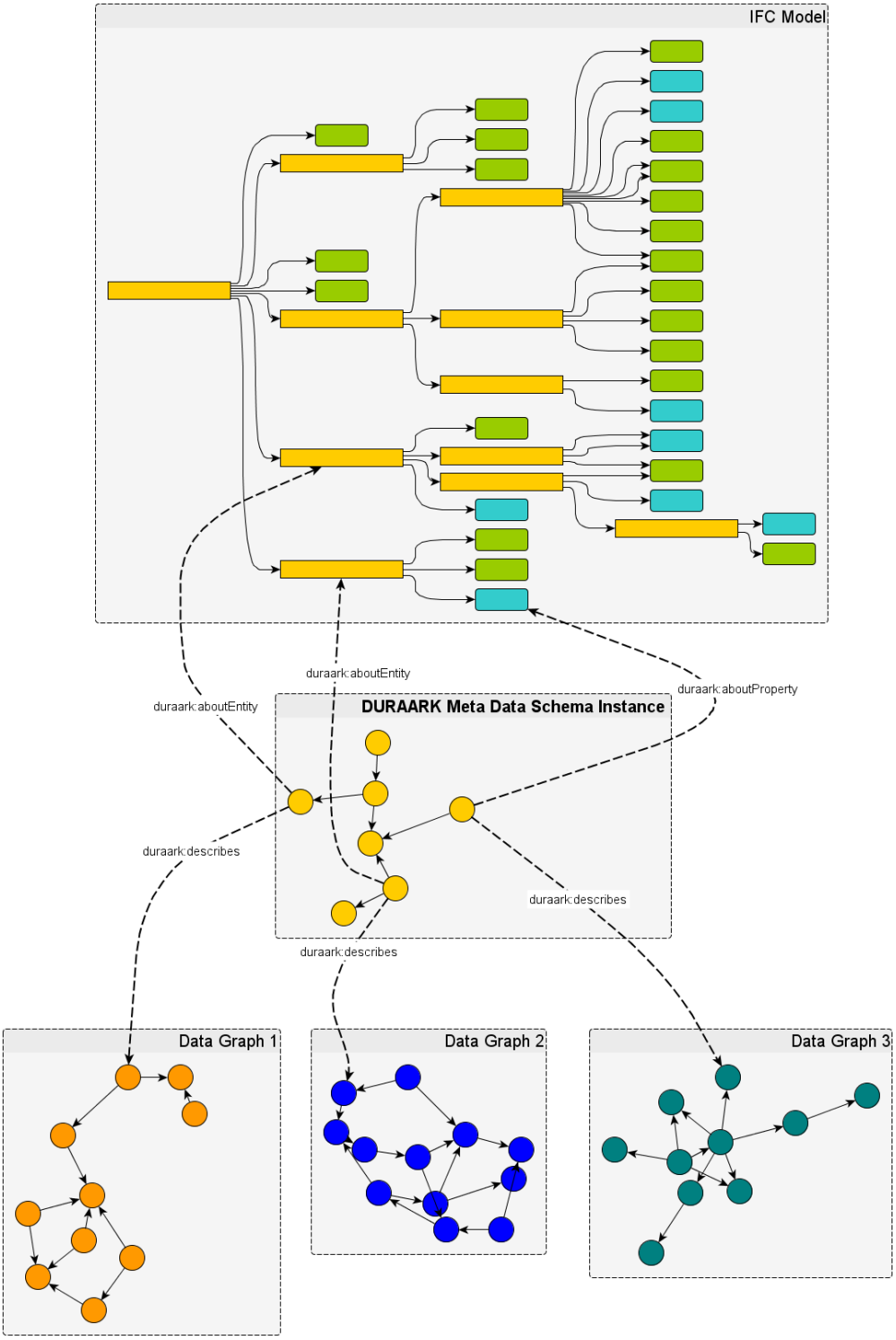

Figure 5: Meta-data schema as a hub for linking external information to IFC models.

Semantic enrichment (technical) of BIM component
This component allows the access to external data sets and the creation of links between entities in these data sets and individual building elements, spaces or the building as a whole. This component focuses on technical engineering data on high levels of detail. Examples of external data sets include classification systems, building regulations, product data bases, and concept repositories such as the buidlingSMART data dictionary bsDD. Snapshots of the content of the linked data will become be added to the SDA. Resources available in the SDA can also be accessed directly. The component is either a stand-alone post-processing tool such as Catenda's 'Bimsync' or integrated in to a modeling CAD application.

- **INPUT** IFC model with of-the-shelf information generated by modeling (CAD) packages.
- **OUTPUT** IFC model enriched with semantic information from external data sources.

5.2.3 Preservation tools

Migration of IFC models for archival purposes ("IFC/A") (pre-ingest)
In order to be reproducible without external dependencies, an IFC model has to prepared for the archival during the ingest phase. While a many issues are either addressed by other software components (the incorporation of point-cloud data) or are outside the scope of the DURAARK project (the concurrent version management of building models during the design and construction phase), a particular challenge lies in the use of external references from within the IFC model itself: Since it is to be expected that an increasing number of properties of single components or whole parts of buildings are specified and defined using external information such as classification systems, building regulations or third-party component libraries in the future, these reference have to be included into the archive in order to be self-containing. As is described in the deliverable of WP3, there are a number of ways, how such external information can be referenced from within and IFC Schema. A principle overview of this mechanism is illustrated in figure 6. A particular challenge with this kind of linked data sets from using external resources is their temporary nature: All of the resources are subject to constant evolution and their

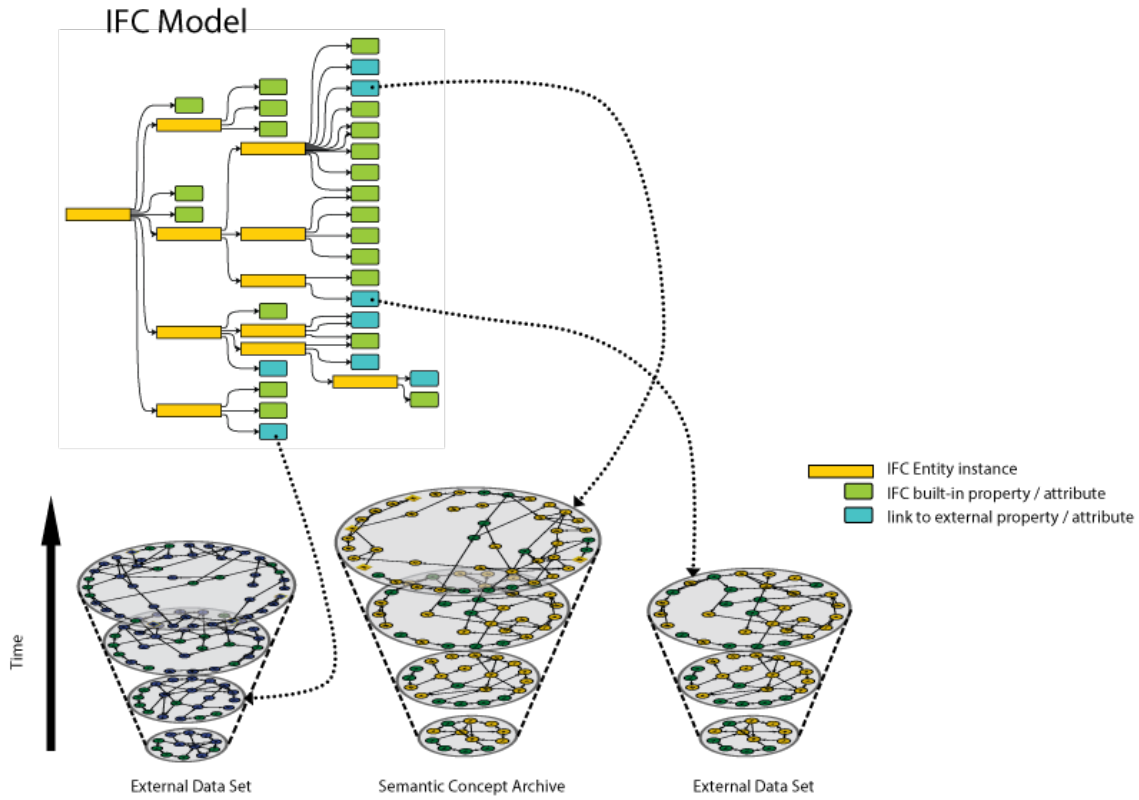

Figure 6: Using the built-in mechanisms to refer to external information.

reference is only valid at a particular point in time. Archiving building models thus means that snapshots of the external data resource will also have to be taken.

- **INPUT** IFC model.
- **OUTPUT** Self-containing IFC/A model.

IFC validation component

This component validates an IFC file for its general schema compliance and the "IFC/A" MVD compliance. Apart from the detection of syntactic and structural errors, it also checks for the presence and validity of information necessary for the archival meta-data generation (authorship, generator tools etc.)

- **INPUT** IFC model.

- **OUTPUT** Validated IFC model and validation log / certificate.

5.3 Semantic meta-data storage

SDA curation component
This component allows the preparation of data sets and their storage in an archive which provides self-containing snapshots of external vocabularies and data sets that can be shared by ingested archival packages.

- **INPUT** URI of external vocabulary or data set.
- **OUTPUT** Self-containing snapshot of external vocabulary and data set.
- **OUTPUT** Update in registry of Semantic Digital Archive.

Interlinking & clustering component
This component is the main processing part for the semantic enrichment. In addition to the meta-data extracted from information explicitly present in the underlying IFC model, a number of aspects relevant to indexing is not available in the model itself but will be retrieved in a semi-automated way. It will deploy mechanisms which identify correlations between data & models and make these links explicit, for instance, to create links between different structure models/BIM which are part of the same building, different models which represent the same structure at different points in time, energy efficiency data and the building it relates to, two structures sharing similar geodata or similar correlated information where explicit links are required in order to ensure consistency. As such, the main purpose of the interlinking and clustering component is to populate the SDA with richer, and more coherent data, i.e. to produce a more coherent *RDF graph of architectural models and related information*. With respect to enrichment and interlinking with external data, such information may include geographical context, related transport and infrastructural information, environmental data, historic information about the structure and its surroundings, public perceptions of structures and their evolutions and similarly relevant information that has to be attached to the mere engineering information covered in the BIM/IFC model by external experts.

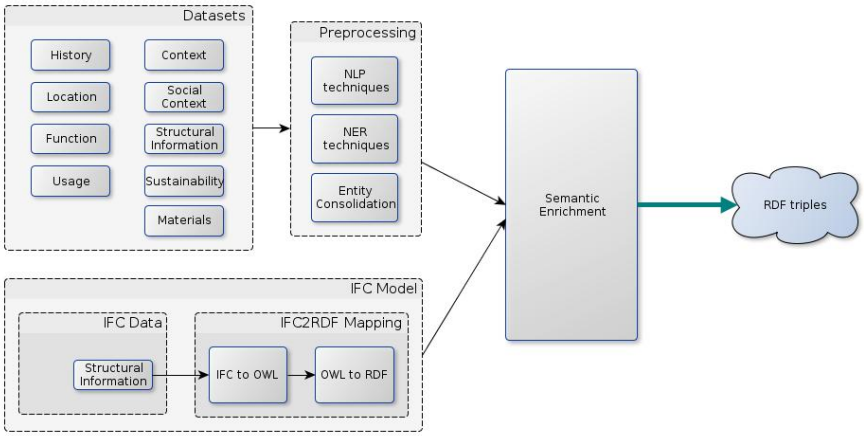

Figure 7: Processing Pipeline of the Semantic Digital Archive.

In the context of the semantic web, a rich number of data sets can be harnessed to add such information. These are currently under investigation as part of WP3 activities. The metadata schema developed in WP 3 allows the creation of links. The software component allows the enrichment of an ingested IFC/A data set with such information. While heterogeneity of datasets (at the schema and instance-level) is a major obstacle, state of the art schema matching[24][14] and entity linking techniques[23] will be deployed to detect and capture links between disparate datasets. In addition, feature-based clustering and machine learning techniques will be deployed to identify links between related models and data.

5.4 Digital preservation system (Rosetta)

Identify file format component

File identification is considered a core process of digital preservation. To treat files in digital preservation processes, it is essential to know what the exact file format is (including version level of the format). Furthermore, it is beneficial if the file format can be referenced through a global unique identifier, such as the PRONOM Unique Identifier (PUID). This component is available in the pre-ingest stage as well as in the digital preservation system itself.

- **INPUT** IFC model.
- **OUTPUT** file format, file format version, PUID (if available).

Risk extraction - check file dependencies component
Risk extraction is similar to technical meta-data extraction. However, whereas technical meta-data extraction captures a lot of information as separate values, risk extraction addresses a single risk at greater detail. Such a risk is the file dependency on embedded or referenced files, such as textures either integrated as a binary large object (bmp, jpg, gif, png) or referenced via a URI. The risk extraction/file dependencies component will analyze the object for any dependencies and give further information about the objects embedded or referenced. This component is available in the pre-ingest stage as well as in the digital preservation system itself.

- **INPUT** IFC model.
- **OUTPUT** number of embedded or referenced objects, availability of embedded or referenced objects, type of embedded or referenced objects.

SIP creation for ingest into digital preservation system component
At the end of the Pre-Ingest processing chain, the object and all relevant descriptive and technical meta-data needs to be packaged as a Submission Information Package (SIP). A SIP follows a specific, predefined file/folder structure and requires a manifest, explaining the order and content of the information object. The acceptable SIP structure is predefined by the digital preservation system that the objects will be ingested into. This component is available in the pre-ingest stage as well as in the digital preservation system itself.

- **INPUT** IFC model(s), accompanying meta-data.
- **OUTPUT** SIP(s).

Submission to digital preservation system component
After successful creation the SIP needs to be passed to the digital preservation system. The submission component will push one or more SIPs to the digital preservation system and log the status of the ingest into the system. Submissions should be planable via jobs.

- **INPUT** One or several SIPs.
- **OUTPUT** Log of SIP status in digital preservation system.

PROBADO - Rosetta connector component

To enable retrievability of the objects from the PROBADO platform, PROBADO needs to hold information about the ID of the objects within Rosetta. The ID will enable PROBADO to generate a link to the Dissemination Information Package (DIP). The ID is ideally captured during the submission job and passed to PROBADO from there.

- **INPUT** SIP.
- **OUTPUT** ID of the object in Rosetta.

6 Conclusion

In this deliverable, the overall software architecture of the DURAARK project has been presented. The constraints and decisions towards this architecture are based on the functional- and non-functional requirements. The corresponding key components have been identified and described.

This document will be complemented with the deliverable D2.2.3 System Architecture & Specification v2.0 in M12, which will give a detailed description of all used interfaces and specification.

References

[1] ASTM International Technical Committee E57. http://www.astm.org/COMMITTEE/E57.htm.

[2] JHOVE - JSTOR/Harvard Object Validation Environment.

[3] libE57: software tools for managing e57 files (ASTM e2807 standard). http://www.libe57.org/.

[4] PREMIS Data Dictionary for Preservation Metadata, version 2.0. Technical report, Mar. 2008.

[5] G. Antoniou and F. van Harmelen. Web ontology language: Owl. In *S. Staab, & R. Studer, Handbook on Ontologies*, pages 67–92, 2004.

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

[7] R. Berndt, I. Blümel, M. Clausen, D. Damm, J. Diet, D. W. Fellner, C. Fremerey, R. Klein, M. Scherer, T. Schreck, I. Sens, V. Thomas, and R. Wessel. The probado project – approach and lessons learned in building a digital library system for heterogeneous non-textual documents. In M. Lalmas and et al., editors, *Research and Advanced Technology for Digital Libraries, 14th European Conference ECDL. Proceedings ECDL 2010*, volume 6273, pages 376–383. Springer, 2010.

[8] R. Berndt, I. Blümel, H. Krottmaier, R. Wessel, and T. Schreck. Demonstration of user interfaces for querying in 3d architectural content in probado 3d. In *Research and Advanced Technology for Digital Libraries*, pages 491–492, 2009.

[9] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[10] M. Böhms, P. Bonsma, M. Bourdeau, and A. S. Kazi. Semantic product modelling and configuration: challenges and opportunities. *ITcon Special Issue Next Generation Construction IT*, 14:507–525, 2009.

[11] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[12] K.-H. H. Cheung, H. R. Frost, M. S. Marshall, E. Prud'hommeaux, M. Samwald, J. Zhao, and A. Paschke. A journey to Semantic Web query federation in the life sciences. *BMC bioinformatics*, 10 Suppl 10(Suppl 10):S10+, 2009.

[13] M. Crawford, R. O'Leary, and J. Cann. *Uniclass: Unified Classification for the Construction Industry*. Riba Publications Limited, 1997.

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

[15] S. Dietze, D. Maynard, E. Demidova, T. Risse, W. Peters, K. Doka, and Y. Stavrakas. Entity extraction and consolidation for social web content preservation. In A. Mitschick, F. Loizides, L. Predoiu, A. Nürnberger, and S. Ross, editors, *2nd International Workshop on Semantic Digital Archives*, volume 912 of *CEUR Workshop Proceedings*, pages 18–29. CEUR-WS.org, 2012.

[16] S. Dietze, S. Sanchez-Alonso, H. Ebner, H. Yu, D. Giordano, I. Marenzi, and B. P. Nunes. Interlinking educational resources and the web of data - a survey of challenges and approaches. *Emerald Program: electronic Library and Information Systems*, 47(1), 2013.

[17] S. Dietze, H. Yu, C. Pedrinaci, D. Liu, and J. Domingue. Smartlink: a web-based editor and search environment for linked services. In *8th Extended Semantic Web Conference (ESWC), Heraklion, Greece*, 2011.

[18] M. Dolenc, P. Katranuschkov, A. Gehre, K. Kurowski, and Z. Turk. The InteliGrid platform for virtual organisations interoperability. *ITcon*, 12:459–477, 2007.

[19] T. Erl. *Service-Oriented Architecture: Concepts, Technology, and Design*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2005.

[20] D. L. Federation. <mets> metadata encoding and transmission standard: Primer and reference manual. version 1.6 revised., 2010.

[21] C. Lima, C. F. Da Silva, C. Le Duc, and A. Zarli. A framework to support interoperability among semantic resources. In *Interoperability of Enterprise Software and Applications*, page 87–98. Springer, 2006.

[22] W. Mazairac and J. Beetz. Towards a framework for a domain specific open query language for building information models. In *Proceedings of the 2012 eg-ice Workshop*, Technische Universität München, Germany, 2012.

[23] B. P. Nunes, S. Dietze, M. Casanova, R. Kawase, B. Fetahu, and W. Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking. In *ESWC 2013 - 10th Extended Semantic Web Conference*, May 2013.

[24] B. Pereira Nunes, A. Mera, M. Antonio Casanova, B. Fetahu, L. A. P. Paes Leme, and S. Dietze. Complex matching of rdf datatype properties. In *In Proceedings of 24th International Conference on Database and Expert Systems Applications*, 2013.

[25] M. Smith. MIT FACADE Project Final Report. Technical report, August 2009.

[26] F. Tolman, M. Böhms, C. Lima, R. van Rees, J. Fleuren, and J. Stephens. eConstruct: expectations, solutions and results. *Electronic Journal Of Information Technology In Construction (ITcon)*, 6:175–197, 2001.