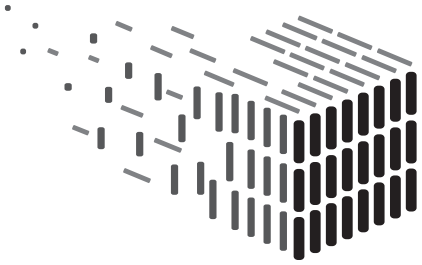


D3.4 Semantic Digital Interlinking and Clustering Prototype

DURAARK

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

Date: 2014-10-31
Version 1.0
Document id. : duraark/2014/D.3.4/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.3.4
Title of document	:	Semantic Digital Interlinking and Clustering Prototype
Deliverable type	:	Software prototype
Contractual date of delivery	:	2014-10-31
Actual date of delivery	:	TBD
Lead beneficiary	:	Catenda
Author (alphabetic)	:	Jakob Beetz <j.beetz@tue.nl> (TUE) Stefan Dietze <dietze@l3s.de> (L3S) Besnik Fetahu <fetahu@l3s.de> (L3S) Ujwal Gadiraju <gadiraju@l3s.de> (L3S) Thomas Krijnen <t.f.krijnen@tue.nl> (TUE)
Responsible editor(s)	:	Jakob Beetz <j.beetz@tue.nl> (TUE)
Quality assessor(s)	:	Sebastian Ochmann <ochmann@cs.uni-bonn.de> Martin Tamke <Martin.Tamke@kadk.dk> Östen Jonsson <osten.jonsson@ldb-centrum.se>
Approval of this deliverable	:	Marco Fisichella Stefan Dietze
Distribution	:	Public
Keywords list	:	Semantic Enrichment, Ontologies, Archival, Preservation

Executive Summary

In this report, a description of the interlinking and clustering mechanisms and prototypes is provided that will be integrated into the DURAARK workbench and the Semantic Digital Archive (SDA) backend. This report will be expanded by D3.6:

1. **D3.4**, delivered in M21, introduces the general concepts and initial prototypes of interlinking and clustering and documents the approaches taken in the prototypical software tools to be implemented as part of the DURAARK system. It also contains results of preliminary experiments carried out in lab-environments which will be augmented and re-evaluated under practical conditions in later stages.

2. **D3.6**, delivered in M30, extends the D3.4 additional practical applications of these prototypical tools, in particular for the focused crawling component. Based on larger scale experiments, the prototypes will be tested and evaluated on additional datasets an by end-users by means of workshops and crowdsourcing with WP7.

In this **D3.4** report three aspects of interlinking and clustering are covered:

1. **Interlinking** of data, knowledge resources and semantically rich metadata of building models including **a software prototype for the creation and validation of links** by end-users

2. **Clustering** of Linked Data concepts that are not yet interlinked including a **a software prototype for the clustering of data sets and vocabularies**

3. **Mashing** of building metadata instances including **a software prototype to mash-up building data with social media**

Table of Contents

1 Introduction

In order to comprehensively preserve building models for future use, two main categories of information have to be captured in a preservation framework like the DURAARK system:

1. The available **raw data** of the building itself. For the DURAARK context the special focus of this information type lies on the explicit, 'hand-crafted' geometric models that are captured in IFC files and as-build information stemming from on-site measurements that are captured in E57 point cloud data. Next to the former two, a wide range of other information types such as text documents, drawings and photographs are of high importance for a comprehensive digital preservation of a building but are only partially considered in the DURAARK context.

2. The **semantically rich metadata** of the building that stem from semantic enrichment of the initially available information and data. As described in earlier DURAARK reports, this semantic enrichment is produced at different stages of the life cycle of a building object, including **pre-ingest** stages during which e.g. the material qualities of individual components, building products used for particular engineering solutions or building code classifications are used to annotate parts of the building or the building as a whole. These annotations are produced by architects, engineers, urban planners and craftsmen of different trade during the planning, construction and operation of the building. An increasing number of such annotations stem from external resources and e.g. refer to classification systems, product catalogues and measurement data residing outside of the the static schema of the IFC format. In order to fully reproduce the whole of a building model, these external types of information have to preserved alongside the IFC and E57 representations of the intellectual entity.

   A second category of metadata is produced **during or post-ingest** and mainly helps downstream applications of archived building data. These applications include scientific analysis of whole buildings, single construction parts and the the role of individual buildings in the evolvement of the city fabric. Often this information consists of curated metadata such as architectural style, ownership and location or aggregated information like number of stories, layout type or use. Such annotations of a building often has to be produced manually by the archiving curator (e.g

architectural style) and will not be provided by the creators of the buildings and building models themselves. In some cases however, the generation of such metadata can be assisted by semi-automatic tools that e.g. that have been implemented in the earlier prototypes of the DURAARK workbench (number of stories, geographic location, level of development etc.). Some of these support tools, have already been developed and described in earlier deliverables of the project.

1.1 Use Cases for interlinking and clustering

The work documented in this report focuses on aspects further assisting the semantic enrichment of the preserved buildings that are focused on **Linked Data**. Typical use cases of such semantic enrichment using Linked Data include:

- The use of **Linked element classification Data** to annotate individual components during planning, construction and documentation stages: "A wall is classified as an external load-bearing wall according to a local building regulation (German: "DIN 276, Kostengruppe 331, Tragende Außenwände"; Dutch: "Nl-SfB 21.2, buitenwanden; constructief"; US: "Omniclass 21-02 20 10 Exterior Walls"). While today such annotations are often defined ad-hoc as mere string values attached to an element, the use of external classification systems such as the buildingSMART Data Dictionary (bSDD)[1] allows the provision of URI links or other unique identifiers for unambiguous referral. For a preservation system, this means that the related classification must be preserved at the time of its use.
  An architectural user or energy consultant of a preservation system might e.g. search for *"all residential buildings with wooden external load-bearing walls in the south of Germany"* in order to compile and assess examples for own projects.

- The use of **Linked product Data**: A particular heat exchanger unit has been used in an office building "Seifert Systems Air-/ air heat exchanger LT 5025-230V; Ambient air circuit 82 cfm @ 50Hz / 88 cfm @ 60Hz; Protection Class NEMA 12; RAL 9018; Order Number 502500011;". Although a few product classification systems such as eCl@ss[2], ETIM[3] or baudataweb[4] are available, almost no product

[1] http://bsdd.buildingsmart.org/
[2] http://www.eclass.de/
[3] http://www.etim-international.com/
[4] http://www.freeclass.eu/

manufacturer relevant to the building industry uses such systems yet. Partially, this is due to the lack of details in the classification systems available (unambiguous definitions of "air exchange rate"[5]) and partially the reasons can be found in the technical obstacles to e.g. embed relevant metadata into micro-formats embedded into the product specification pages.
A user from the Facility Management domain might for instance, use a preservation system to query for *"all chiller products X that have last been maintained 2 years ago"* in order to identify potential optimization in energy use.

- The use of **Linked building classification Data** to annotate a building as a whole: In the urban area of Tel Aviv, buildings have been surveyed and groups of buildings have been classified as belonging to the "Bauhaus" or "International Style" architectural style along with the construction dates and architects. Such classifications today can be accomplished using e.g. the Getty Arts and Architecture Thesaurus (AAT), Union List of Artist Names (ULAN) and Getty Thesaurus of Geographic Names (TGN). Although not yet frequently used, such classifications are on the rise and the underlying vocabularies are mature and stable enough to expect its proliferation in the near future.
  An art historian might use a linked preservation system to query for *"the provenance of Tel Aviv architectural styles in modern European design schools"* in order to achieve a better understanding of cultural influences.

- The use of **Linked perception Data**: The acceptance and popularity of existing or even historic buildings can play a crucial role in the success of planning and design or refurbishment of nearby or otherwise related buildings. Is the steel-glass construction in the middle of a medieval town generally appreciated or rejected? Do form factors play a role in this perception? Are courtyard layouts of mosques valued in cool climates? The relation of buildings to feelings, sentiment and biases is difficult to capture, yet would be a valuable resource for a number of scenarios in planning and building. Research shows the usefulness of such endeavours, and the semantic annotation of buildings e.g. in archives is a necessary prerequisite for such results.
  A planner might consult a preservation system by querying it for *"all buildings in a*

[5] In the above example, the US-American air-exchange rate is provided as 82 cfm (cubic feet per minute) while on the German product pages it is stated as 140 m^3/h (cubic meter per hour)

city X built between 1970 and 1990 with a negative connotation that are currently unoccupied and not protected by a heritage institution" in order to identify potential improvements of urban areas.
A building operator, architect or other stakeholder might consult a preservation system by querying it for, *the perception score of a building* to keep track of the evolving feelings of people towards a building and its surroundings, help to ensure adequate maintenance and trigger retrofit scenarios where required.

- The use of **Linked technical Data** can improve the development of new technologies in the building and construction domain. Such annotation of ingested building data include e.g. the software versions of the tools used for its production (ArchiCad 13, Revit 2014), the schema versions of the populations (IFC 2x3, IFC 4), measurement metadata of point clouds (devices, weather conditions) and registrations of point clouds in explicit IFC models. Although not referring to *semantic* metadata like the use cased provided before, the availability of technical metadata according to the vocabularies provided in earlier DURAARK deliverables as well as the ability to query them across individual archives is highly appreciated by the construction informatics community.
A researcher developing new object recognition methods for point cloud data might query a preservation system for *"all residential buildings of at least 3 stories represented as IFC 4 with a registered point cloud dataset of at least 40 measurement points using device X in non-rainy conditions including colored images of at least 10000x5000 px"* in order to acquire test data sets for a new algorithm.

For all use cases stated above that are further elaborated in earlier deliverables and mission statements of the DURAARK project, a prerequisite is the existence of readily available, preferably open vocabularies and data sets that allow end-users to annotate ingested buildings. An initial overview of such data sets as DBPedia and vocabularies such as bSDD can be found in D3.1. However, currently such are Linked Data sets are very limited and are not yet used on a frequent basis. As part of the DURAARK vision to provide a preservation system for architectural data, a central hub for such data sets and vocabularies is the **Semantic Digital Archive** that also preserves the states and snapshots during its evolvement. This SDA is not only meant as a storage mechanism for already enriched building models, but should also spur the use of semantic enrichments by serving as a domain-specific registry for building related data sets and vocabularies.

An essential part of such a registry is to gather, group, classify and align existing Linked Data and provide a sustainable set of tools to allow the addition and evolvement of this registry and archive with future (versions of) Linked data sets.
For the creation of such registries in the SDA, this report covers three main aspects of **Linking** (section 3) and **Clustering** (section 4) of data. One of the envisaged applications of practical interlinking and clustering is documented in section 5 that illustrates how such interlinked metadata can be used for **Mashing** building data with social media.

2 Software Prototype: Access and Usage

For this Deliverable 3.4 a number of prototypical software modules have been implemented which will be further described in the respective sections in the remainder of this document.

2.1 Manual Interlinking Prototype

- **Live installation:** The interlinking prototype (3) can be found on http://www.duraark.eu/interlink which is will be the permanent link to the latest working version of the tool. Currently it is physically located on http://bw-dssv19.bwk.tue.nl/interlink/. The interlinking prototype will also be included in later revisions of the overall DURAARK workbench.
- **User Manual:** The Use of the software if further detailed in section 3.2.1.
- **Source Code:** The source code can be found https://github.com/DURAARK/interlink

2.2 Automated Clustering Prototype

While the current prototype performs crawling and relevance computation as a backend process, a live installation, integrated into the DURAARK workbench along with a user manual will be provided as a part of the D3.6. The crawler will be directly integrated into the DURAARK semantic enrichment processes and will help populating the SDA & SDO developed in WP3.

- **Source Code:** The source code developed to realize the objectives within the automated clustering prototype (4) can be found at https://github.com/DURAARK/focussed_crawler/.

2.3 Automated Data Linking Prototype (Social & Semantic Web)

- **Source Code:** The source code developed to realize the objectives within the automated data extraction and linking prototype in Section 5 can be found at https://github.com/DURAARK/building_perception/.

- **Dataset:** More information about the dataset we considered, the ground truths established for each building type and experimental results can be found at http://data-observatory.org/building-perception/.
- **SPARQL End-point:** We provide the following SPARQL end-point at which our data can be queried; http://meco.l3s.uni-hannover.de:8829/sparql.
- **Additional Visualization:** We use MapBox's TileMill[6] in order to build interactive maps showcasing our findings and representing the architectural structures in our dataset. This link[7] presents an interactive map depicting the popularity of airports around the world, i.e. how these airports are perceived. The size of the circle representing each airport signifies the magnitude of perception. By hovering over the airport circles on the map, a user can harness information regarding the airport, including its corresponding normalized popularity score. Further details are provided when a user clicks on an airport.

[6] https://www.mapbox.com/tilemill/
[7] https://a.tiles.mapbox.com/v3/ujwal07.4qu84cxr/page.html?secure=1#2/0/0

3 Linking

In the DURAARK context, especially in the SDO (Semantic Digital Observatory) module it is important to address the heterogeneity of data that is registered (and preserved) in the SDA (Semantic Digital Archive) in a form of a structured registry. In order to allow end-users such as archivists, curators and other stakeholders to enrich building models with particular semantic information and metadata (as e.g. illustrated in the introduction), the need arises to group, classify and relate available datasets and vocabularies with respect to their potential use ("*here*, opinions about buildings can be gathered." "*This* vocabulary allows the classification of building materials per component." "Vendor-independent descriptions of building products are described using *this* classification."). In this particular case, we focus on *clustering* of the data by measuring different relatedness measures which exploit the RDF representation of the data, as well as their unstructured (e.g. textual content) representation. This naturally leads to the second point of consideration, namely *inter-linking* of resources as part of the SDA. In the following we describe the process of *clustering* and *inter-linking* of resources residing in the DURAARK SDA.

3.1 Means of Linking

Linking refers to the process relating resources to each other. Examples of such resources include concepts in an ontology ("Draw Bridge"), items in a classification (see examples in 1.1), words in a dictionary or individual objects such as the description of the Eiffel Tower in different data sets. In the case of Linked Data captured in RDF such relations can be captured by (i) inserting assertions as RDF statements into one of the Data Sets (or both) or (ii) storing the links in separate logical units such as files or named graphs in a triple store. Such relations between different resources can then be combined e.g. in ad-hoc queries ("Which facts are known about the Eiffel Tower when combining data sets X, Y, Z?", "What are the properties of the concept "Retaining Wall" that should be preserved in a Facility Management Context in the Netherlands?") or used to synthesize specialized vocabularies or data sets.

3.1.1 Alignment

The process of linking two data sets is often also referred to as **Alignment**. In general, two different kinds of alignments can be distinguished:

1. **Schema alignment** refers to linking two modelling vocabularies[8] and their populations such as RDFS, Dublin Core or SKOS to find similar relations (e.g. `rdfs:Class` to `skos:Concept`). It allows the creation of large number of links between two datasets ("everything being a type of 'class' in vocabulary A) can be transformed into being a type of 'concept' in vocabulary B). For vocabularies 'in the wild' the context of the DURAARK project[9], such schema alignments however are seldom used since popular modelling languages are widely reused or have already been mapped. An exception are sometimes property mappings for example of partonomic relations that are often custom made in engineering ontologies.

2. **Named Entity alignment** refers to the creation of mappings between instantiations or individuals of schemas. As an example, the 80.000 concepts in the bSDD are instances of the `owl:Class` construct that have many other properties attached to them. These can be semi-automatically aligned to e.g. the millions of YAGO[10] concepts by looking at the labels attached to the concept or other similarity measures further described in section 5.

3.1.2 Semantics and syntax of links

In ontology-, knowledge-engineering and data modelling a wide spectrum of possible relations between two items can be modelled. The meaning of these relations however is often difficult to define precisely in such a way that they can be communicated, shared and reused unambiguously. From the many different relations between data items, information types or knowledge facets on various granularity we are focusing on a few selected ones for the use of semantic enrichment in the DURAARK context. These include

Similarity and sameness In a Linked Data context one of the most frequent relations among resources has the intention of stating "A is *'the same as'* or *'similar to'* B" including granular distinctions of weight such as *'almost the same'*, *'exactly the same'* or *'roughly comparable to'*. The differences between these notions are often subtle, yet have a considerable impact depending on their interpretation and use. Next to such simple similarity relations it is often desired to express contextual

[8] often referred to as the "Terminology Box" or "T-Box" in knowledge engineering

[9] see also earlier deliverables D3.1 and D3.2

[10] a mash-up created and curated from combining Wikipedia WordNet and GeoNames datasets, see https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

relatedness such as "similar with regard to shape" (e.g. the glass-steel entrance pavilion to the Louvre by architect I.M. Pei and the Cheops Pyramid). If such similarity with regard to a particular aspect or facet is modelled with false relations it might lead to wrong conclusions (e.g. the Louvre is located in Kairo)
Technically, similarity and sameness can be in expressed using relationship predicates from a number of vocabularies: `owl:sameAs`, `owl:equivalentClass` and `skos:exactMatch` are among the most popular ones. Their use depends on the resources that should be related and semantics that should be expressed by these links.
An extensive discussion of the issues related to similarity and sameness relations as well as a proposal of a more fine-grained similarity ontology can be found in [13]

Specialization To indicate that two resources A and B are related by each other in terms of *'being more specific than'*, *'being a special case of'*, *'belonging to the broader category of'* or simply *'being a type of'*. Often such relations are referred to as 'taxonomic' or subsumption relations and a number of technical means in RDF vocabularies can be used to express them. As with similarity the different properties such as `rdfs:subClassOf`, `skos:broader`, `skos:narrower` a great deal of attention has to be paid to their proper use and applications (see also [6]).

Containment To structure ontological facets of vocabularies ("buildings according to form" vs. "builings accoring to function") or to group individuals ("the category of French Art Nouveau buildings") different modelling vocabularies offer collection and grouping mechanisms such as `skos:member` of a `skos:Collection`, `owl:UnionOf` or DBPedias categories.

3.2 Manual interlinking prototype implementation

As already illustrated in the introduction in section 1, currently only a very limited number of comprehensive and detailed building related vocabularies for the building and construction sector exist. However, for the various semantic enrichment use cases also described in the introduction, the availability of such vocabularies is highly desirable. Many aligning and linking tools available today are either targeted for structured vocabularies and ontologies with a limited amount of resources, or do not scale well. To close this gap,

the vocabulary registry of the SDA should support the growth of such vocabularies with a specific focus on building related data sets.
In this subsection a prototypical tool is introduced that helps to manually validate pre-computed links based on clustering mechanisms and tools introduced in section 4.

3.2.1 Hands-on evaluation of the prototype

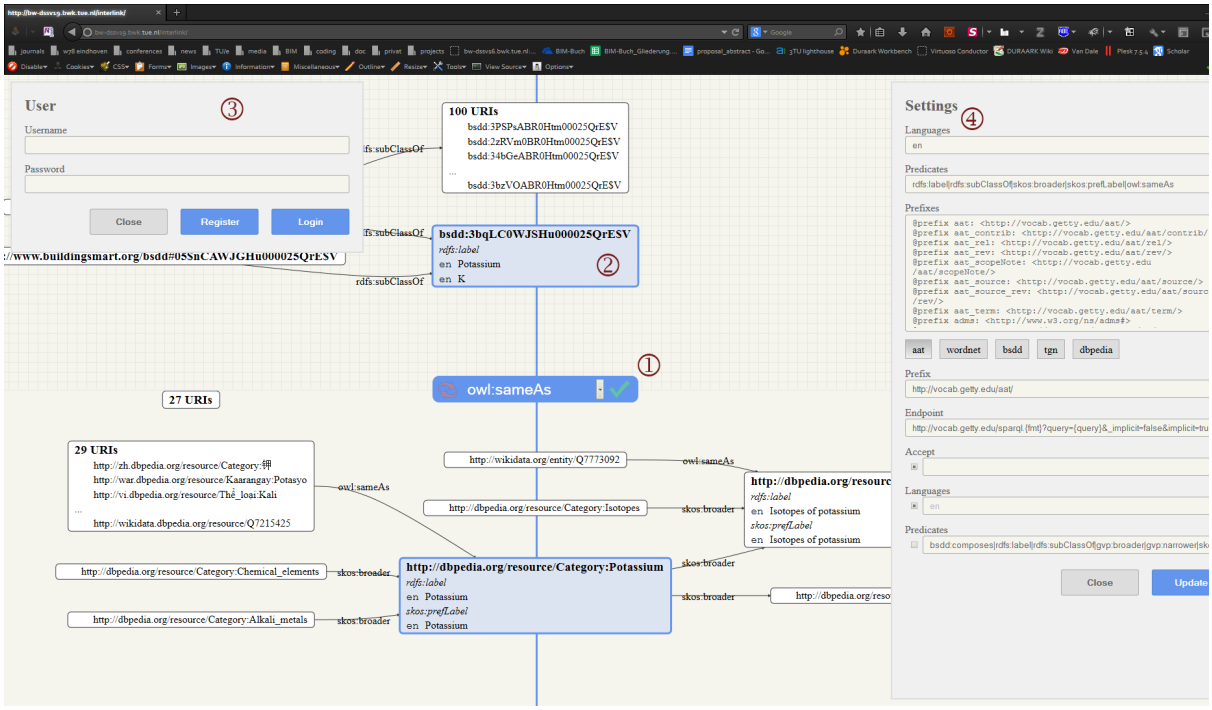


Figure 1: Screenshot of the Manual Interlinking Prototype showing two concepts being linked. 1)relation widget with vertical alignment axis. 2)Node widget showing the labels of the selected node. 3) Register/Login dialog. 4)Configuration Dialog

A working version of the prototype implementation of the manual interlinking prototype can be found on http://bw-dssv19.bwk.tue.nl/interlink/ or the permanent URL http://durraark.eu/interlink which will redirect users to the latest working version. In later stages, the tool will be integrated into the DURAARK workbench where it be a part of the SDA operation and maintenance tools.
For this initial prototype release, A random link of pre-configured data endpoints (bSDD, AAT, WordNet, DBPedia, Semantic Concdepts) for which semi-automated links have been created with separate tools will be presented to the user. Upon clicking on the check-mark

on the relations widget (1 on the UI screenshot in figure 1) a VoL record will be generated and saved into the SDA data store[11]. Before confirming a link, the neighbourhood of the node can be explored by single-clicking on the bold URL-labels of the widget (2 on the UI screenshot in figure 1). If another relation seems more appropriate a click into an empty area of the widget will center the respective node on the vertical alignment axis.
Users might also choose to register and login (3 in figure 1) or configure different endpoints in the configuration dialogue (4 in figure 1). Reloadeing the page or pressing the reload icon on the relation widget selects another preconfigured link.
The documented source code can be found in the SVN repositories of the DURAARK project and on https://github.com/DURAARK/interlink

3.2.2 Use case, target audience and design rationale

In order to make the SDA registry component useful for the semantic enrichment of building models, well-suited tools for the large-scale creation of such mappings and alignments is desirable. To accelerate the creation of larger clusters of interlinked vocabularies specific to the needs of the building industry and the semantic enrichment requirements for long term archiving, a possible approach lies in the 'crowd-sourcing' of such mappings. The target audience of such efforts are domain experts and practitioners with high confidence in the building and construction domain but limited technical resources and know-how.

3.2.3 Requirements

From the boundary conditions stated above, a number of requirements for such a tool can be compiled:

1. **Installation and technical requirements** The interlinking tool must be accessible with a low technical threshold. Typical mapping and alignment front-ends often require the installation e.g. a Java Runtime with granted security privileges or are plug-ins on top of highly complex and sophisticated environments (Protege) targeted at scientific audiences.The tools should allow minimal the use with minimal set-up.

2. **Scalability** The manual interlinking tool should be able be usable with large vocabularies and datasets such as bSDD, AAT or YAGO. It should also enable

[11] currently implemented as a separate database that will eventually be moved into the main SDA repository

the concurrent work of many (tens or even hundreds) users than. In the long run, re-enforcement, audit and peer-reviewing mechanisms should be integrated that would allow the emergence of communities cross-validating each others' work.

3. **Extendibility** Vocabularies and datasets for building and construction are constantly evolving: New resources are added, existing vocabularies change their content or technical interfaces and some disappear. The tool should thus enable the interaction

4. **Usability** In order to allow an informed decision making concerning the linking between any two given nodes, it is often crucial to take the context into consideration. To allow this, users should be enabled to explore the surroundings of a given node in the respective graph. Currently only a very limited amount of readily-available software tools allow a coherent, exploitative and graphical interaction with large graphs.

3.2.4 Conceptual approaches of the prototype

A crowd-sourced manual interlinking platform has been added to the DURAARK framework. It allows end-users to pairwise, manually link concepts that reside in two ontologies or other datasets or to evaluate alignments that have been generated using semantic clustering and interlinking methods described in earlier sections.
In order to do so, the user is presented with the two terms from these ontologies and the option to link them. The context of the terms can be browsed interactively, to enable the user to assess that the two terms share the same meaning by means of looking at their relating terms, or to make sure that a broader or narrower term isn't a better match for the link. An overview of the interface is provided in figure 1.
It is important to distinguish between the various different potential semantic meanings of the links. The meaning of the link will have profound implications for example on the inference rules that can be applied when a user subsequently queries the body of linked ontologies. The links that are available for selection are listed in listing 1. Of these links `owl:sameAs` (for individuals / instances) and `owl:equivalentClass` (for `owl:Class`es) have the strongest meaning, whereas `rdfs:seeAlso` carries little semantics at all. The former states that the two URIs share the same identity (are interchangeable) and therefore all relations to one of the terms is said to be applicable to both (see also earlier section 3.1.2)

```
owl:equivalentClass
owl:sameAs
skos:closeMatch
skos:exactMatch
skos:narrowMatch
skos:relatedMatch
skos:broadMatch
rdfs:seeAlso
```

Listing 1: Examples of the (configurable) ralationship types that can be used to link given nodes in a graph

The link chosen to be most suitable by the user is recorded in a Vocabulary of Links (VoL) [12] structure to be able to annotate the created link with provenance data. An example of such a VoL structure is provided in listing 2. This facilitates the preservation of provenance information such as ownership which is recorded as a part of the interlinking process. As a practical application, filters can be applied that e.g. only process links generated by certain users or organisations, rank and weight links, remove vandalism or anonymous links etc.

3.2.5 User interface

The interface is built as a web-based component of the DURAARK framework. As such it is easily distributed to individuals that want to take part in the crowd-sourcing initiative. The core User Interface (UI) part consists of two horizontally aligned layered graphs. Layered graph drawing is commonly used for Directed Acyclic Graphs (DAGs) but does not necessarily rule out the possibility of rendering cycles. Some of the ontologies considered to be useful for the DURAARK context conform to a tree for their most prominent concepts. For example, the concepts in the Getty Art and Architecture Thesaurus (AAT) span a tree, because it complies to SKOS. The layered graph drawing presents a structured visual appearance to the user and enables users to easily ascend or descend the tree of broader and narrower terms by clicking the respective widgets. Because the graph layout algorithm will typically enforce that broader terms will reside on a lower rank than the narrower terms. This structured presentation helps users to asses whether broader or narrower terms are a more viable target for the link. This is an advantage over force-directed graph layout algorithms which are employed by most other graph visualization and navigation applications. These do not distribute nodes over various ranks and therefore cannot

[12] http://data.linkededucation.org/vol/

present a hierarchical ontology in a meaningful way. Furthermore force-directed graph layout algorithms tend to flicker, as they oscillate towards an equilibrium, and typically do not cope well with nodes that vary greatly in size nor with directed graphs.
The list of potential matches, as used by the prototype, can be acquired by using state-of-the-art ontology alignment tools, such as the SILK workbench. The resulting listing of URI pairs can be fed into the SQL database by means of a conversion script that is supplied along with the prototype application.

3.2.6 Technical implementation details

Architecture The client-side UI is built on top of a Javascript library called Dagre [13], which itself makes use of D3 [14], a very common platform for data visualization using modern web standards. In addition the code is structured using module loading by means of RequireJS [15]. This enforces modular and re-usable code. By default Dagre sorts nodes within ranks to minimize the amount of edge crossings. This is disabled to maintain the internal ordering of the nodes when nodes are expanded as the user traverses the related terms.

Persistence The implementation provides some elementary user management so that administrator users can filter contributions by users or institutions. This is implemented using a trivial server application on a LAMP (Linux Apache MySQL PHP) stack. In addition, this server stores a list of potential matches, which are pairs of URIs from different or the same ontologies. Through a JSON REST interface the server allows the retrieval of a random entry from this list. After the user reviews a link, an intermediate entry is stored on the server. This constitutes the pair of uris along with the match semantics and the responsible user. This intermediate result can be expanded into an RDF turtle representation in which each unique pair of uris in the result set is written as a vol:Link. If multiple entries exist for the same pair, the most prevalent link semantics will be selected as the vol:hasType attribute and each non-anonymous contributer will be listed as a dc-term:creator for this link. As such, this RDF result can be fed back into a triple store to query the body of linked ontologies.

[13] https://github.com/cpettitt/dagre
[14] http://d3js.org/
[15] http://requirejs.org/

Configuration A small PHP-based proxy server is provided to work around Cross-Origin Resource Sharing requirements enforced in modern browsers. The same origin policy dictates that either Javascript resources should be on the same domain, or a HTTP response header should be added to the responsive explicitly allowing the client to make use of this resource. At the time of writing not all SPARQL endpoints provided functionality to configure such behavior. Since a server-side proxy does not enforce the Same Origin policy dictated by the browsers, this requirement is circumvented by rerouting the SPARQL queries, that the UI emits to explore the context of the aligned terms, through this proxy server.

```
@prefix owl: <http://www.w3.org/2002/07/owl#>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
@prefix skos: <http://www.w3.org/2004/02/skos/core#>
@prefix vol: <http://purl.org/vol/ns#>
@prefix dc-term: <http://purl.org/dc/terms/>

<interlink> a vol:LinkingMethod ;
rdfs:comment "A web-based tool for the manual verification of automatic ontology
    alignments"@en ;
dcterms:creator <http://www.ddss.nl/Eindhoven/staff/Jakob.Beetz> ;
dcterms:creator <http://www.ddss.nl/Eindhoven/staff/Thomas.Krijnen> .

<link_9ead93628a4d7fecfcc8cbe9b8d0d631> a vol:Link ;
vol:linksResource <bsdd:0kVxwAPKWHuO00025QrE$V> ;
vol:linksResource <http://vocab.getty.edu/aat/300051254> ;
vol:usedLinkingMethod <interlink> ;
vol:hasType owl:sameAs ;
dc-term:creator <krijnen> .
```

Listing 2: Example of a VoL record with provenance information created by the manual interlinking prototype: An owl:sameAs link has been created between

4 Clustering

4.1 Means of Clustering

Clustering is a process which is in most cases used to address heterogeneity of data, by grouping related resources based on attributes that describe them. In this case, we deal with resources coming from linked datasets (as part of the SDA), which in themselves are structured. However, a great deal of attributes from such resources are in textual form, hence, unstructured.
Such textual attributes (literal datatypes) and as well (object properties) are used as main features to represent resources in SDA:

1. **literal values** (e.g. textual abstract describing a particular resource)
2. **object values** (e.g. resource URI)

Using the defined features for representing a resource in SDA, in the following sections we describe the similarity measures and the clustering techniques. However, before going into any detail of the similarity and clustering techniques, we provide an overview of the data as part of the SDA and how we generate the ground-truth for evaluating the clustering techniques described in this report.

4.1.1 Dataset Analysis & Ground-truth Construction

In this section, we provide a thorough analysis of the datasets as part of the SDA. This serves the purpose of reasoning for the appropriateness of used similarity measures and clustering techniques.
The analysis is carried using one of the largest existing Web of Data dataset, namely the BTC dataset (Billion Triples Challenge)[16]. The dataset, as the name suggest, consists of billions of triples resulting in more than 300GB of uncompressed data. The dataset has been compiled from other linked datasets such as DBPedia or Freebase. Even though these data sets are of a generic nature, they contain a considerable amount of resource relevant to buildings and architecture: For example, the Freebase data sets currently contains about 135.000 topics and 400.000 facts in the 'Architecture' Domain. The choice of using this data set despite the fact that is only of limited immediate interest and relevance to the preservation of building is based mainly on two reasons: 1) A large dataset allows the

[16] http://km.aifb.kit.edu/projects/btc-2012/

evaluation of the tool with respect to performance and scalability. 2) Building-specific data sets of this size are not yet available.
Considering the two features used to describe resources in SDA, in Table 1 we show the frequency of specific literal datatype properties and the corresponding average length in terms of number of words.

datatype property	frequency	avg. length
http://www.w3.org/2000/01/rdf-schema#label	91521315	23.48
http://www.w3.org/2000/01/rdf-schema#comment	74898887	28.71
http://www.w3.org/2004/02/skos/core#prefLabel	61773105	32.22
http://xmlns.com/foaf/0.1/knows	18773659	13.81
http://xmlns.com/foaf/0.1/nick	17516745	9.33
http://xmlns.com/foaf/0.1/member_name	11538878	11.41
http://xmlns.com/foaf/0.1/tagLine	11538877	22.79
http://xmlns.com/foaf/0.1/name	9235251	12.69
http://xmlns.com/foaf/0.1/primaryTopic	8513234	12
http://xmlns.com/foaf/0.1/homepage	8244952	43.72
http://dbpedia.org/ontology/abstract	7948551	270.17
http://usefulinc.com/ns/doap#name	3772606	30.33
http://rdfs.org/sioc/ns#follows	3713750	30.15
http://purl.org/dc/terms/date	3697989	12.55
http://purl.org/linked-data/sdmx/2009/measure#obsValue	3625256	4.63
http://purl.org/dc/elements/1.1/title	3605629	14.58
http://xmlns.com/foaf/0.1/maker	3142730	13.21
http://xmlns.com/foaf/0.1/accountName	2818424	9.06
http://xmlns.com/foaf/0.1/interest	2810540	11.33
http://xmlns.com/foaf/0.1/topic	2696607	33.06

Table 1: Average length in terms of number of words for the different datatype properties and the frequency in the BTC-12 dataset.

From Table 1 we can see that a majority of resources in the BTC-12 dataset, have a literal datatype property like `rdfs:label` with an average length of 23 words. Based on this analysis, such datatypes are used in later stages to describe the particular resources, which in turn are used to measure the similarity (usually when considering lexical similarity) between any pair of resources for a given clustering approach.
Apart from the literal datatype properties, a very important feature of linked datasets are object datatypes. These represent one of the basic principles of linked data, namely interlinking resources on the web. Such links between resources can have different semantics, i.e. `rdfs:seeAlso` or `skos:related`, convey the relatedness of a given pair of resources.

object datatype	frequency
http://www.w3.org/2000/01/rdf-schema#seeAlso	24153844
http://xmlns.com/foaf/0.1/knows	18773600
http://www.w3.org/2002/07/owl#sameAs	18665783
http://purl.org/dc/terms/subject	11607974
http://xmlns.com/foaf/0.1/primaryTopic	8513172
http://xmlns.com/foaf/0.1/homepage	8244754
http://dbpedia.org/ontology/wikiPageExternalLink	4904972
http://dbpedia.org/ontology/wikiPageRedirects	4243200
http://rdfs.org/sioc/ns#follows	3713737
http://xmlns.com/foaf/0.1/page	3469910
http://ontologycentral.com/2009/01/eurostat/ns#geo	3447580
http://xmlns.com/foaf/0.1/maker	3142551
http://xmlns.com/foaf/0.1/account	2816644
http://xmlns.com/foaf/0.1/accountProfilePage	2816610
http://rdfs.org/sioc/ns#account_of	2810831
http://xmlns.com/foaf/0.1/interest	2810513
http://xmlns.com/foaf/0.1/topic	2694875
http://purl.org/ontology/mo/performer	1890499
http://rdf.freebase.com/ns/common.topic.article	1848120
http://xmlns.com/foaf/0.1/isPrimaryTopicOf	1783520

Table 2: Frequency of the top-20 object datatype properties in the BTC-12 dataset.

However, as seen in Table 2, there are much more object datatype properties, which can be further used to measure similarity between a pair of resources, without having any strict semantics as the ones mentioned above(`rdfs:seeAlso` or `skos:related`). Such object datatypes, as we will describe later are used in similarity metrics that exploit the graph nature of linked datasets to measure relatedness of resources.

However, in general such clustering approaches come at a price of scalability (comparing labels requires trillions of operations). Therefore, as shown in Figure 2, based on the distribution of resources to the particular types, we focus our clustering and inter-linking efforts for those types that have a higher number of resources (the labels of resource types are omitted to improve the clarity of the plot). In addition, we ignore the types that are at the tail of this distribution. This serves us for the purpose of narrowing down our clustering and inter-linking only to a subset of dataset residing in the SDA.

Finally, for the evaluation of the clustering techniques we use as a ground-truth dataset for the resources which are inter-linked through object datatype properties that have clear semantics like *relatedness*, *equivalence* etc. (see Table 2 for details). The ground-truth

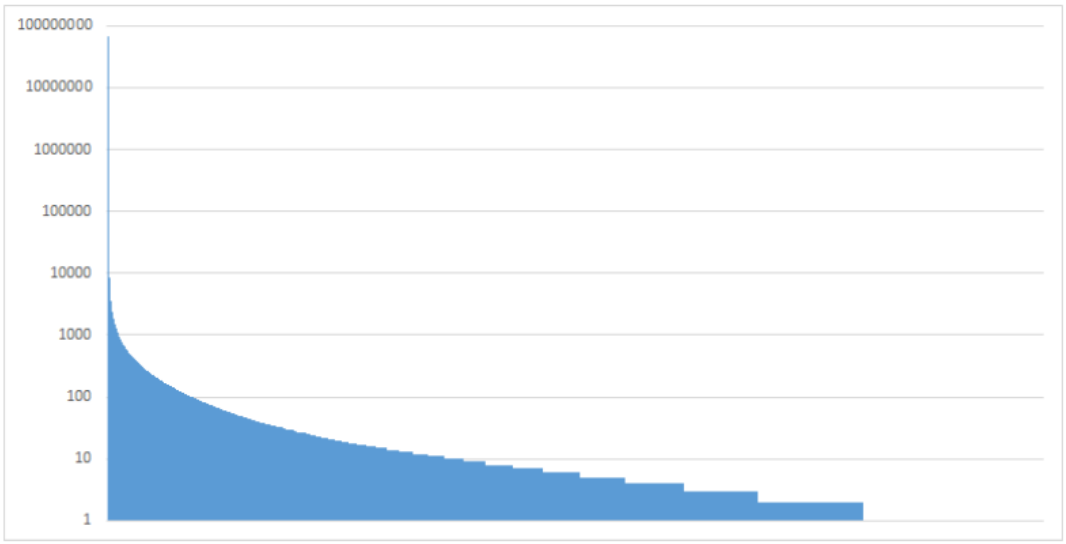


Figure 2: Resource distribution across the different resource types in the BTC-12 dataset.

dataset, extracted from the BTC-12 dataset, serves to set thresholds for the different similarity metrics used in our clustering process.

4.1.2 Similarity Metrics

Based on the defined attributes that are used to describe resources in the SDA, we use two types of similarity metrics: i) lexical and ii) structural.

Lexical similarity In the first group of lexical similarity metrics, we consider the following.

- **Cosine similarity**, it measures the similarity between a pair of resources, respectively their vector representation of weighted terms based on the `tf-idf` weighting scheme. As described earlier, to generate the vector representation of a resource, we rely on the datatype properties in Table 1.
- **Jaccard distance**, is another lexical similarity metric, which measure the overlap of terms from a pair of resources. The term vector representation of a resource is generated as in the previous similarity metric.

Structural metrics In the second group of structural similarity metrics, we consider the following.

- **Shortest path distance**, here we measure the shortest path between a pair of resources, based on the Dijkastra's shortest path algorithm. Such a path consists of links between resources that are connected through the object datatype properties (see Table 2).
- **Katz index**, specifically the modified version of the Katz index by Nunes et al. [27]. It measures the similarity of a pair of resources based on the number of paths between them, by penalizing exponentially longer paths in favour of shorter paths.
- **SimRank** [16], is usually used on scenarios where the context of the pair of resources is important. In details, it gives the contextual similarity of resources, by measuring how similar are the other resources that refer to the given pair of resources.

The above similarity metrics present an initial set of metrics that cover both aspects of unstructured and structured datasets in the SDA. However, this list will be revised and enriched with more metrics as is deemed useful during tests and evaluation.

4.1.3 Clustering Approaches

In this section, we enumerate the clustering techniques that are used. We consider multiple clustering approaches in order to measure the performance of such techniques in two aspects: *clustering accuracy* and *clustering scalability*.

The first approach under consideration is the**k–means** clustering algorithm. It is one of the most widely used clustering approaches. As a similarity metric we can use both the *lexical* and *structural* metrics. Furthermore, it can be applied on larger datasets, such as our case in DURAARK. However, one of the disadvantages of using such an approach is determining the number of appropriate clusters given a dataset. In most cases, this is done using multiple initialisations of **k–means** with a varying number of clusters.

The second approach we consider is **Spectral Clustering** [42]. The procedure is as follows: Given a set of n resources that we want to cluster together, compute the adjacency matrix between the given set of resources. In our case, the adjacency between a pair of resources is given by one of the similarity metrics. On the computed adjacency matrix $\mathbf{A}$ we perform singular value decomposition, from which in turn we extract the eigenvectors and eigenvalues (see for details [42]). From the computed eigenvectors and eigenvalues,

using the $k-$means algorithm we are able to cluster the resources with high accuracy. The performance of the $k-$means algorithm is heavily influenced by the number of clusters and analysed dimensions from the eigenvector space. We determine the number of clusters relying on a heuristic [42], which is commonly adopted in the case of spectral clustering. The number of clusters and analysed dimensions from the eigenvector space is determined by the first spike we detect in the eigenvalue distribution. That is, whenever for two consecutive eigenvalues their difference is significantly higher than from the previous eigenvalues), is the point at which the best clustering accuracy is achieved.

4.2 Initial Experiments

As a first step towards realizing the aforementioned approaches for clustering resources and assessing their relevance to the information need (provided by means of a 'seed-list', an initial set of information items relevant e.g. to architecture), we conducted a number of initial experiments.

4.2.1 Seed-list

The goal of focused crawling in this work is to find related and high quality entities for a specific set of seeds. The seed list collectively represents the information intent.
For our initial experiments, we consider a seed-list consisting of 30 skyscrapers presented in the table 3. The diversity in the location of the skyscrapers represents a broad scope in the information intent. In the following subsections, we present results pertaining to this seed-list.
Our subsequent experiments in the future will consider the following aspects, in order to arrive at optimal configurations for crawls as well as, clustering and relevance assessment measures:

- Varying scope in the intent of a crawl.
- Diversity within seed entities.
- Size of the seed-list.

4.2.2 Crawling

We first carried out some experiments in order to determine the suitability of two existing approaches for crawling the Web of Data, in particular Linked Data: (i) LDSpider: An

Named Entity	Disambiguated DBPedia URI
Tower 185	dbp:Tower_185
Fountain Place	dbp:Fountain_Place
1540 Broadway	dbp:1540_Broadway
US Bank Tower	dbp:US_Bank_Tower
Die Pyramide	dbp:Die_Pyramide
One America Plaza	dbp:One_America_Plaza
777 Tower	dbp:777_Tower
Mellon Bank Center	dbp:Mellon_Bank_Center
One Worldwide Plaza	dbp:One_Worldwide_Plaza
Comcast Center	dbp:Comcast_Center_(Philadelphia)
Museum Tower	dbp:Museum_Tower_(Dallas)
Two Prudential Plaza	dbp:Two_Prudential_Plaza
Enterprise Plaza	dbp:Enterprise_Plaza
Thanksgiving Tower	dbp:Thanksgiving_Tower
Aon Center	dbp:Aon_Center_(Chicago)
Tower Life Building	dbp:Tower_Life_Building
900 North Michigan	dbp:900_North_Michigan
Chrysler Building	dbp:Chrysler_Building
One Museum Park	dbp:One_Museum_Park
Philadelphia City Hall	dbp:Philadelphia_City_Hall
Two California Plaza	dbp:Two_California_Plaza
Manchester Grand Hyatt Hotel	dbp:Manchester_Grand_Hyatt_Hotel
Messeturm	dbp:Messeturm
Trammell Crow Center	dbp:Trammell_Crow_Center
Wells Fargo Plaza	dbp:Wells_Fargo_Plaza_(Phoenix)
Comerica Bank Tower	dbp:Comerica_Bank_Tower
Empire State Building	dbp:Empire_State_Building
The Trump Building	dbp:40_Wall_Street
Willis Tower	dbp:Willis_Tower
Woolworth Building	dbp:Woolworth_Building
Fox Plaza	dbp:Fox_Plaza_(Los_Angeles)

Table 3: Seed-list consisting of skyscrapers and corresponding disambiguated DBPedia URIs.

open-source crawling framework for the LD [15], and (ii) LDCrawler: An iterative Linked Dataset crawler for Preservation [17].

Results from our benchmarking experiments showed that the performance of the LDSpider is better. Therefore, we use the breadth-first approach with the LDSpider for further

[17] Crawl Me Maybe: Iterative Linked Dataset Preservation. Besnik Fetahu, Ujwal Gadiraju, and Stefan Dietze. In Proceedings of the 13th International Semantic Web Conference, 2014.

experiments. The runtime performance of the LDSpider with varying number of hops are presented in the Table 4. We find that increasing the number of hops to beyond 2 deters the crawl due to the exponential growth in the corresponding graph size.

Number of Hops	Time Used (s)	Size(M)
0	3.5	1.1
1	279.2	60.9
2	27226.1	3943.6

Table 4: Run-time performance of the LDSpider.

4.2.3 Term-based Relevance Computation

In order to generate the vector representation of resources, in the form of weighted terms based on the `tf-idf` weighting scheme as described earlier, we use the textual datatype properties of `rdfs:comment` and `dbp:abstract`.
We use the Cosine similarity (CS) and Jaccard Distance (JD) metrics to compute the pairwise similarity between candidate entities and seed entities. Table 5 presents the term-based pairwise similarity of the candidate entities in our experimental crawl using the LDSpider and the seed-list consisting of the skyscrapers.

No. of Hops	No. of Candidate Entities	Time Used (s)	Avg. CS	Avg. JD
1	2023	2.4	0.043937	0.016857
2	560720	147.6	0.017211	0.014327

Table 5: Average term-based pairwise relevance of candidate entities with respect to seed entities.

We note that with the increasing number of hops, more candidates are retrieved with a cost of lower relevance. In order to establish thresholds for relevant candidate entities we need to adopt empirical methods. Table 6 presents the most relevant candidate entities with respect to the seed entities as per the Cosine Similarity metric, while Table 7 presents the same with respect to the Jaccard Distance metric.

4.2.4 Graph-based Relevance Computation

A graph is built with the candidate entities as nodes and the relationships between them as edges. We use Vertex Cosine Similarity between the graph representation of the

Relevant Candidate Entity	Cosine Similarity
dbp:Renaissance_Tower	0.1566
dbp:Nicholas_Tower	0.1169
dbp:30_Hudson_Street	0.1143
dbp:Alcide_de_Gasperi_Building	0.0945
dbp:Nebraska_Township,_Livingston_County,_Illinois	0.0897
dbp:Hotel_des_Indes_(Batavia)	0.0606
dbp:Recordando_Otra_Vez	0.0582
dbp:Eric_Bloom	0.0461
dbp:Eastern_State_Penitentiary	0.0407
dbp:Inez_Courtney	0.0353

Table 6: Most relevant candidate entities as measured using Cosine Similarity between candidate and seed entity pairs.

Relevant Candidate Entity	Jaccard Similarity
dbp:Tower_Life_Building	0.0806
dbp:Williamsburgh_Savings_Bank_Tower	0.0405
dbp:Federal_Reserve_Bank_of_Chicago	0.0339
dbp:Tacony_Music_Hall	0.0319
dbp:Child:_Music_for_the_Christmas_Season	0.0252
dbp:Hollenden_Hotel	0.022
dbp:Brooklyn_Chinese-American_Association	0.0174
dbp:1995_Gallery_Furniture_Championships	0.0083
dbp:Walter_Cramer	0.0076
dbp:Ed_Kemmer	0.006

Table 7: Most relevant candidate entities as measured using Jaccard Similarity between candidate and seed entity pairs.

candidate entities with respect to the seed entities. Vertex Cosine Similarity gives the cosine similarity between vertices of a graph. Table 8 presents the most relevant candidate entities with respect to the seed entities as per the Vertex Cosine Similarity metric.
As a next step, we will conduct further experiments with other graph-based similarity measures such as shortest path distance, and Katz index as mentioned earlier.

Relevant Candidate Entity	Vertex Cosine Similarity
dbp:Brown_Building_(Manhattan)	0.3429
dbp:Bob_Mader	0.1562
dbp:Phoenix_Corporate_Center	0.1538
dbp:John_Craig_Freeman	0.1515
dbp:Brittany_Brooks	0.1515
dbp:Lemmie_Miller	0.125
dbp:Lamesa,_Texas	0.125
dbp:5th_Reserve_Officers'_Training_Corps_Brigade	0.125
dbp:USS_Jason_(AR-8)	0.0938
dbp:Neta_Snook	0.0938
dbp:William_Stephen_Devery	0.0625
dbp:Her_Double_Life	0.0312
dbp:Henry_Mitchell_(oceanographer)	0.0312

Table 8: Most relevant candidate entities as measured using Vertex Cosine Similarity between candidate and seed entities in their graph representation.

5 Mashing Building Data with Social and Semantic Web Data

Knowledge about the reception of architectural structures is crucial for building owners, architects or urban planners. The evolution of such perception of buildings can also be useful for large building corporations, to understand the impact of structures over time. Yet obtaining such information has been a challenging and costly activity. With the advent of the Web, a vast amount of structured and unstructured data describing architectural structures has become available publicly. This includes 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). Hence, first mining (i) the popularity of buildings from the social Web and (ii) then correlating such rankings with certain features of buildings, can provide an efficient method to identify successful architectural patterns. We propose an approach to rank buildings through the automated mining of Flickr metadata. By further correlating such rankings with building properties described in Linked Data we are able to identify popular patterns for particular building types (airports, bridges, churches, halls, and skyscrapers). Our approach combines crowdsourcing with Web mining techniques to establish influential factors as well as ground truths to evaluate our rankings. Our extensive

experimental results depict that methods tailored to specific structure types allow an accurate measurement of their public perception.

5.1 Background

Urban planning and architecture encompass the requirement to assess the popularity or perception of built structures (and their evolution) over time. This aids to better understand the impact of a structure, identify needs for restructuring or to draw conclusions useful for the entire field, for instance, about successful architectural patterns and features. Thus, information about how people think about a building that they use or see, or how they feel about it, could prove to be invaluable information for architects, urban planners, designers, building operators, and policy makers alike. For example, keeping track of the evolving feelings of people towards a building and its surroundings can help to ensure adequate maintenance and trigger retrofit scenarios where required. On the other hand, armed with prior knowledge of specific features that are well-perceived by the public, builders and designers can make better-informed design choices and predict the impact of building projects.

Until now there has been limited research in tackling the problem of ranking architectural structures based on their associated perception. So far, obtaining feedback about the perception of buildings has been a challenging and costly, yet important activity for stakeholders. Gathering such data historically required a significant amount of manual labour. With the advent of the Web, a substantial amount of data has become available publicly. This data provides information about the perception and use of buildings, for instance through social media. The social Web provides a multitude of channels, such as Twitter, Flickr, Foursquare, etc. for users to voice their opinions about situations and contexts in which they are in, often involving particular buildings. This provides a rich source for deriving information about the popularity and perception of certain structures of different types, such as airports, churches, bridges, and so forth. The Web also contains structured information about particular building features, for example, size, architectural style, built date, etc. of certain buildings through public Linked Data. Here in particular, reference datasets such as Freebase[18] or DBpedia[19] offer useful structured data describing a wide range of architectural structures.

[18] http://www.freebase.com/
[19] http://dbpedia.org/

The perception of an architectural structure itself has historically been studied to be a combination of the aesthetic as well as functionality aspects of the structure [39, 40]. The impact of such buildings of varying types on the built environment, as well as how these buildings are perceived, thus varies. For example, intuitively we can say that in case of churches, the appearance plays a vital role in the emotions induced amongst people. However, in case of airports or railway stations, the functionality aspects such as the efficiency or the accessibility may play a more significant role. This suggests that the impact of particular *influence factors* differs significantly between different *building types*. We introduce a processing pipeline and experiments which mine the Social Web in order to measure (rank) popularity of buildings. We exploit the Web of data to correlate building rankings with corresponding features, in order to enable identification of statistically more popular architectural patterns. Through this work, the important contributions are as follows.

- We present a method for ranking architectural structures.
- An approach to gain further insights into the perception of architectural structures, by bridging the gap between the Social and Semantic Web (correlation of structure features with facts from the Social Web).
- Influential factors and ground truths for ranking architectural structures as well as automatic models for generating accurate rankings.
- The generated data itself (consisting of architectural structures and their perception) has been exposed as public Linked Data, and in addition published through an interactive visualization in the form of a conjunct map.

5.2 Related Literature

Little work has been done in trying to understand the aesthetic appeal of an architectural structure and its affect on the surrounding environment, in the context of exploiting Web data. Research has however, established the fact that the architectural structures play an important role in influencing the built environment and consequently the well-being of a community. Leyden et al. show that the design and conditions of cities are strongly associated with the happiness of residents in 10 different urban areas[22]. Lathia et al. reflect on community well-being from urban mobility patterns [21]. Bill Hillier introduced

space syntax, a science-based, human-focused approach that investigates relationships between spatial layout and a range of social, economic and environmental phenomena[14]. These phenomena include patterns of movement, awareness and interaction; density, land use and land value; urban growth and societal differentiation; safety and crime distribution. In his seminal book [1], Christopher Alexander presents notions on the contextual nature of building perceptions. The author introduces the concept of '*a nameless quality*' that we should seek to include in all buildings, and discusses patterns at the level of architectural design constructs (for example, low roof-lines, east-facing windows, and so on). In another work, Alexander et al. introduce patterns as timeless entities that present problems with respect to the architectural design of buildings and towns, and then offer a solution to each problem[2]. In our work, we complement their observation that 'most of the wonderful places of the world were not made by architects but by the people'. We therefore attempt to mine architectural patterns based on various influence factors depending on the type of structure.

There has been a large amount of research concerning employing the wisdom of the crowds, to solve tasks which require a large amount of human input or computation. Such works have also spanned across various domains. The authors of [28] suggest making crowdsourcing an integral part of the workflow for Galleries, Libraries, Archives and Museums (GLAMs). The author of [33], crowdsources perceptions of beauty, quiet and happiness across the city of London by using Google Street View images. The authors of [34], combine gamification and crowdsourcing in order to build a recognizability map of the city. The authors of [12] identify *nichesourcing* to optimize the result of human-based computation for some tasks. They show that nichesourcing combines the strengths of the crowd with those of experts in the relevant field.

There has been a fair amount of research work in the domain of sentiment extraction and analysis from web data sources. Das et al. develop a methodology for extracting small investor sentiment from stock message boards [11]. Bollen et al. explored sentiment analysis on a Twitter dataset, and by exploiting a six-dimensional mood vector, they show that events in the social, political, cultural and economic realms have a significant affect on the various dimensions of public mood [5]. Chen et al. explore the problem of automatic extraction of sentiment expressions from tweets, and recognize the usefulness of assessing the target-dependent polarity of each sentiment expression in a tweet, instead of associating sentiment with an entire tweet [8]. The authors of [44] take first steps towards exploring the problem of entity-ranking in a large set of heterogeneous entities. The

authors of [5] present a system that assigns polarity score for each entity in a large corpus of News and blogs, through a phase of sentiment identification followed by sentiment aggregation. Kennedy et al. show that Flickr tags and other metadata can be used to enhance and improve our understanding of the world [18]. The authors of [38] study the connection between sentiment of Flickr images expressed in the corresponding metadata and their visual content, while the authors of [37] explore the influence of sentiment expressed in YouTube comments on the ratings for these comments using SentiWordNet thesaurus, a lexical WordNet-based compilation with sentiment annotations.

In the context of ranking *architectural structures*; over the last decade and more, there has been a considerable amount of research done with an aim towards determining the efficiency or sustainability of a building, and comparing buildings on criteria pertaining to these features[9, 4, 36]. Roulet et al designed a multi-criteria rating methodology for buildings with the purpose of ranking or rating office buildings and retrofit scenarios of the same building according to an extended list of parameters[35]. Similar works have focused on the Indoor Environment Quality of different buildings as a means of comparison and/or ranking buildings. In order to design appropriate evaluation and rating methodologies for buildings, we need to take into consideration a number of characteristics. In general, many parameters and criteria are considered to access the buildings by each of these methodologies. The criteria may include visual and acoustic comfort, cost and energy efficiency, impact on the environment, perceived health and so on. Energy efficiency however, is considered to be the main factor in almost all recent building rating schemes[20]. In [43] Yang et al describe a method of identifying and weighting indicators for assessing the energy efficiency of residential buildings in China. Lombard et al. analyse available information pertaining to the consumption of energy in buildings, and particularly related to heating, ventilation and air conditioning systems[32]. They address the adequacy of the available necessary information and delve into the main building types. They draw comparisons between buildings of different countries and focus on commercial buildings. The case of offices is analysed in greater detail in this work by Lombard et al. Apart from energy efficiency and the other characteristics already mentioned, buildings are also rated and compared based on their safety provisions. Most often this has to do with fire-safety measures [41, 19, 26]. These works however, only consider the functional aspects of architectural structures and fall short with respect to gauging the aesthetic elements.

5.3 Problem Definition

In the following sections, we formalize the notions that we aim to address. Through our work, we first aim to establish automated methods to compare and consequently rank architectural structures of varying types. Finally, by correlating the rankings with structured data from DBpedia about building characteristics, we demonstrate how successful architectural patterns can be automatically identified.

As a first step towards achieving this, we attempt to find answers to the question, '*How does a building make one feel?*'. We can formalize these notions as follows: We define *Influential Factors* as the aspects that influence the perception of an architectural structure. Let $B = \{b_1, b_2, ..., b_i, ...b_n\}$ be the set of buildings or structures, and $T = \{t_1, t_2, ..., t_k, ...t_z\}$ be the set of building types (for example, churches, halls, skyscrapers). Given the set B of type t_k, we want to determine an optimal ranked subset, F, of *influential factors* which play a vital role in influencing building perception among people.

We thereby want to analyse the varying influence of the factors in the set, $F = \{f_1, f_2, ..., f_i, ...f_m\}$ on different building types in T. Let $Profile(b_i)$ be the building profile consisting of web data relevant to each building b_i. We formulate the perception of a building b_i, as the normalized sum of sentiments expressed towards the building, with respect to the various *influential factors*.

$$Perception(b_i) = \frac{1}{|F|} \sum_{j=1}^{m} Senti^{f_j}(Profile(b_i))$$

$Senti^{f_j}$ represents the sentiment score determined by using the influential factor f_j for the building b_i.

Next, we present methods to automatically rank buildings of a particular type t_k based on the emotions that are invoked by the buildings among people, i.e. according to the perception of the buildings, $Perception(b_i)$. By exploiting the ranking of architectural structures thus generated, and correlating them with characteristics $C = \{c_1, c_2, ..., c_i, ...c_x\}$ of the buildings (extracted from DBpedia), we draw insights into architectural patterns. The characteristics in C map to DBpedia properties for the respective building type t_k. We define an *architectural pattern* as a linear combination of mappings from building characteristics in C (e.g. *architectural style*) to a particular value or value range (e.g. *gothic*). We aim to identify successful architectural patterns, where 'success' is proportional to the positive perception of a structure.

5.4 Approach

In this section we explain our approach to rank architectural structures and mine successful architectural patterns.

5.4.1 Overview

We follow a threefold approach in order to rank structures based on their perception and consequently find patterns of well-perceived architectural structures. (i) First, we identify the *Influence Factors*. (ii) Next, we rank structures by crowdsourcing their popularity, in order to form the ground truth. In addition we use automated methods for sentiment analysis and ranking. (iii) Finally, we correlate the influence factors with related structured data from DBpedia in order to identify well-perceived patterns for architectural structures. We define a well-perceived *pattern* as one that results in a high positive $Perception(b_i)$ value, for any structure b_i (for example, churches with a particular architectural style or skyscrapers with a height between x-y metres).
Figure 3 depicts our approach to combine crowdsourcing and Web mining methods, tailored to the type of architectural structures.

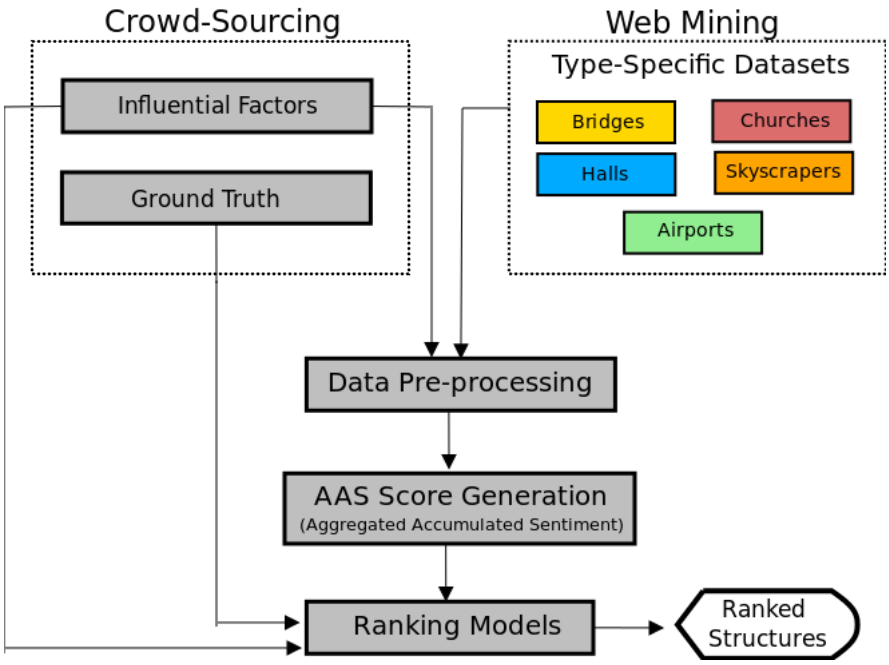


Figure 3: Pipeline reflecting our tailored approach for ranking architectural structures.

5.5 Crowdsourcing Influential Factors

Recent research works in the field of Neuroscience [29, 30], reliably suggest that neurophysiological correlates of building perception successfully reflect aspects of an architectural rule system that adjust the appropriateness of style and content. They show that people subconsciously rank buildings that they see, between the categories of either high-ranking ('sublime') or low-ranking ('low') buildings. However, what exactly makes a building likeable or prominent remains unanswered. Size could be an influential factor. At the same time, it is not sound to suggest that architects or builders should design and build only big structures. For instance, a small hall may invoke more sublime feelings while a huge kennel may not. This indicates that there are additional factors that influence building perception. In order to determine such factors, we employ Crowdsourcing.
An initial survey was conducted with a primary focus on the expert community of architects, builders and designers in order to determine influential factors. The survey administered 32 questions spanning over the background of the participants and their feelings about certain buildings, of different types (*bridges*, *churches*, *skyscrapers*, *halls* and *airports*). In order to create and host the survey, we used LimeService[20]. Within a two-day window, we received 42 responses from the expert community. The survey itself can be found at http://data-observatory.org/duraark_survey/. The important influential factors that surfaced from the responses of the survey are presented below.

- The **history** associated with a building was identified to be an influential factor, in terms of its affect on the people. There is a semblance of reverence towards historically significant buildings, and more often than not, they have a positive affect on people.
- The immediate **surroundings** or the **built environment** of a building play a vital role in how the building itself is perceived. We observe that there is variance in what is perceived to be positive, between buildings that fit well into their surroundings and those that stand out.
- The **materials** used in the structure also influence the perception of the building.
- The **size** of a building influences its recognizability and/or visibility. This goes on to influence how the building is perceived.

[20] http://www.limeservice.com/

- **Personal experiences** involving a building play a key role in influencing one's feelings towards a building.
- The **level of detail**, which is an inherent part of a building's structure is an important aspect to consider. We observe varying perceptions of intricate and complex work in the structure of a building. Some people are highly receptive of great craftsmanship, while others prefer more minimalist art work. This includes decoration and ornaments.

These influential factors pertain to the building types *bridges*, *churches*, *skyscrapers* and *halls*. However, we realize that when it comes to airports, people tend to acknowledge the importance of functional aspects of the buildings. By accounting for the functionality aspects that surfaced through crowdsourcing, and referring to Skytrax[21] (a UK-based consultancy that runs an airline and airport review and ranking site), we have arrived at the following list of influential factors for airports.

- **Ease of access** to the airport (car, public transport connections, parking, etc.)
- **Efficiency** of movement/processing inside the airport (to and from gates/terminals, security, length of required paths/time from check-in to gate etc.)
- **General design** and **appearance** (comfort, ambience, natural light, views)
- **Choice/availability** of shops, cafes, restaurants, etc.
- Seating/ resting/ relaxing /entertainment **facilities** in the airport
- Support for other **miscellaneous facilities** (like ATMs, disabled access, airline lounges, telephone access, washrooms, showers, etc.)
- **Size** of the airport

5.6 Crowdsourcing Ground Truth

We deployed surveys for each of the building types, in order to establish the ground truth in each case, using LimeService[22] and CrowdFlower[23]. Respondents of the survey were

[21] http://www.airlinequality.com/
[22] http://www.limeservice.com/en/
[23] http://crowdflower.com/

presented with the buildings of the corresponding type (see Section 4.1 for dataset), and asked to rate them on a 5-point Likert scale from *Strongly Like* to *Strongly Dislike*, if they had been to the structure or seen it in person. Hence, respondents react to their experience or their impression of such an experience. Apart from this, they were also asked to indicate the degree to which each of the influential factors determined from the initial survey (described earlier), played a role in their decision. These results were collected in another Likert scale from *Strong Influence* to *No Influence*. Table 9 shows the number of unique participants or workers that contributed to the survey(s) for the different building types. A *response* is considered to be the set of answers pertaining to the corresponding questions of a building b_i of type t_k. The surveys for the *bridges* and *churches* were deployed on LimeService and shared through online social networks in order to trigger responses, while the surveys for the *airports*, *halls* and *skyscrapers* were deployed on the CrowdFlower platform with a monetary incentive for workers' responses. Table 9 reflects this variation in the number of participants as well as responses due to the type of platform used to obtain responses.

In order to maintain the integrity of drawing a comparison between results attained from these different platforms, we adjust for demographics, age and gender factors to avoid a bias of any kind. In order to ensure that the workers provide valid responses devoid of any deception with ulterior motives, we intersperse the questions in the survey with test-questions that can help us detect bots or other malicious workers aiming to make quick money. By doing so, we easily separate *trusted responses* from the *untrusted responses*. On the CrowdFlower platform there is a provision to create such test-questions, collectively called the *Gold Standard*. Since we utilize our personal and social networks to trigger responses for the surveys hosted on LimeService, we notice that the responses we receive are all trustworthy. This can be explained due to the fact that the workers here are either directly related to the administrators of the surveys or related through a reliable network of friends. Moreover, the fact that there is no monetary incentive nor any other form of explicit incentive implies that the workers provide responses without ulterior motives. This is reflected in Table 9, where we observe no *untrusted responses* (UResponses) for Bridges and Halls, the structures for which surveys were hosted on LimeService.

In addition, to prevent further bias in our crowdsourced surveys, we refrain from using images with filters or those which are edited to enhance the object in the image. We therefore use corresponding images obtained from Wikimedia Commons[24] that only include

[24] http://commons.wikimedia.org/wiki/Main_Page

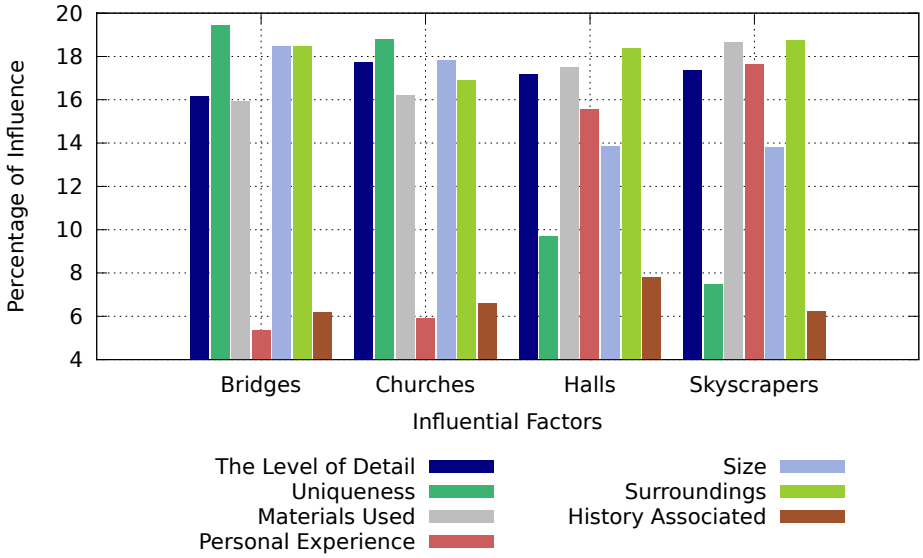


Figure 4: Comparison of Influential Factors for Varying Building Types.

representational excerpts of the surroundings of the structure and devoid of any special touch-ups. We also draw from the findings of the authors in [10], that an image's resolution and physical dimensions affect humans' aesthetic perception of it. The authors verify that generally, higher resolution images are perceived as aesthetically better than lower resolution images. We thereby present images of all the architectural structures in equal resolutions.

By accumulating the responses for each building b_i of type t_k and factoring in the scores from the 5-point Likert scale, we arrive at *normalized popularity-scores* for each building b_i of type t_k. The *normalized popularity-scores* are a reflection of the responses from the workers on the Likert scale, with respect to how they perceive the corresponding structures. We rank the buildings within each building type based on these scores, and adopt these rankings as the ground truth.

The Table 10 presents the various *influence factors* pertaining to *airports* and their corresponding influence in building perception. We observe that all the factors are almost equally influential. Interestingly, the aesthetic factor, '*general design and appearance*' is marginally more influential in building perception than the other functionality related aspects.

The chart in Figure 4 shows a comparison between different building types with respect to the different influential factors. We observe that the '*uniqueness of a building*' plays a

Table 9: Trusted, Untrusted Responses from LimeService and CrowdFlower.

Building Type	# TResponses	# UResponses	# Participants
Airports	5,012	1,441	1,301
Bridges	1,357	0	76
Churches	2,085	0	79
Halls	2,880	641	1,664
Skyscrapers	7,166	370	4,276

significant role in case of *bridges* and *churches*, while it is less influential in the perception of *halls* and *skyscrapers*.

An observation that is common to all the building types is the significant influence of the '*surrounding built environment*' in the perception of a building. This reiterates the mutual influence of a building on its built environment and vice-versa. Essentially, this means that as an extension, one can explore the correlation between a building, and other indices like '*well-being of a community*' or '*the happiness index*', by means of the impact a building(s) has on its built environment.

Similarly, the influence of the *materials used* and the '*level of detail*' are significant across all the building types we consider. The '*size*' of a building, goes a long way in influencing its perception in case of *bridges* and *churches* as opposed to the relatively lower influence in case of *halls* and *skyscrapers*. Personal experiences of people with respect to *halls* and *skyscrapers* seem to influence their perception of the buildings significantly more than *bridges* and *churches*. Finally, the '*history associated*' with a building plays a less influential role towards its perception. We believe this indicates that on average people are either not aware of the historic importance and bearing of most architectural structures, or that their understanding of the historic bearing does not affect their perception of the corresponding structures more significantly.

We found that the influence factors behave similarly for *bridges* and *churches* as opposed to *halls* and *skyscrapers*. Further investigation is required to empirically explain this observation.

5.7 Ranking models using building perception

As shown in Figure 5, we propose to collect data pertaining to each of the buildings b_i in the set B. We create building profiles $Profile(b_i)$ for each of the buildings b_i in the dataset by merging the textual metadata from relevant Flickr images (title, description and comments) into a single representational unit, for each image pertaining to each building.

Table 10: Influential Factors for Airports.

Influential Factor	Influence in Perception
Availability of shops, cafes, etc.	13.26%
Ease of access to the airport	14.58%
Efficiency of movement/processing	14.58%
General design & appearance	15.27%
Relaxing/Entertainment facilities	14.41%
Size of the airport	14.58%
Support for other miscellaneous facilities	14.03%

This data will be used to generate feature vectors corresponding to each building b_i in the list. We will finally exploit different ranking models in order to rank the buildings. In this section we present different ranking models for buildings, based on perception-related data extracted from the metadata of relevant Flickr images.

5.7.1 Sentic Feature Vectors

In order to extract the emotions from the Flickr metadata, we use the National Research Council Emotion Lexicon (EmoLex)[24]. EmoLex is a large lexicon of words annotated with the associated emotions via the means of crowdsourcing[25]. We use this term-based matching technique, which considers that there are 8 main emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) apart from the positive and negative polarity associated with the words, to come up with a significant part of our feature vectors. We define these features, which are a result of employing EmoLex as *sentic features*. Moving further, we adopt a similar approach as in [31]; we create a profile for each building, $Profile(b_i)$, consisting of all the metadata from Flickr images relevant to the building. Then, by using EmoLex we generate a sentic feature vector that represents the various dimensions of emotions contained in the profile of each building. This means that the components of the resultant vector portray each of the 8 emotions elicited by the profile for each building. These 8 components add up to 1 and each of them is a value ranging between 0 and 1. Apart from the 8 emotions, the polarity (positive and negative) features add up to 1 as well.

In addition, we assume that the normalized number of *favorites* for each building and the normalized number of *comments* for each building (accumulated from the metadata of Flickr images relevant to the building) depict the interest of the people towards the building to some extent. This follows our intuition that the *favorites* indicate an approval

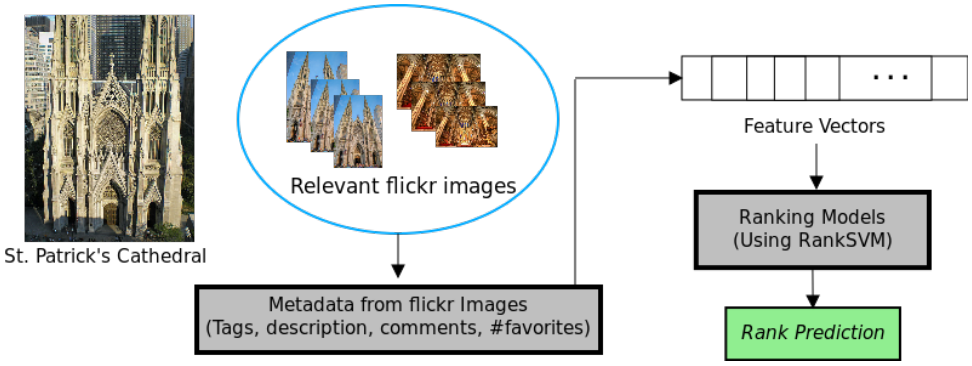


Figure 5: An example illustrating our approach towards the automated ranking of structures.

of the buildings in the images, and can thereby be used as a significant feature to rank buildings automatically. The number of *comments* can also show the interest generated by the building in the picture.
In the ranking models we employ, we follow the steps presented below.

- Using EmoLex we compute the feature vectors for each of the buildings.
- We divide these feature vectors corresponding to all the buildings, into two sets (80%-20%), one for training the model and the other for testing the predictions of the learned model.
- We use Rank SVM to learn a model that can help to automatically rank the buildings based on their corresponding *associated emotions*, since it has emerged as one of the standard pairwise ranking algorithms[17, 7].
- We create 10 splits in order to reasonably gauge the performance of the model from 10 rounds of learning (training) and consequent predictions (testing).
- In order to evaluate the performance of the ranking models, we use the Normalized Discounted Cumulative Gain (NDCG) metric. NDCG is a commonly used metric to judge the performance of an algorithm on training data and to compare the performance with other machine-learned ranking algorithms. Furthermore, by computing NDCG at different levels we can gain insight into the quality of the trained models.

$$NDCG(T,k) = \frac{1}{|T|}\sum_{j=1}^{|Q|} Z_k \sum_{m=1}^{k} \frac{2^{rel_m}-1}{log(1+m)}$$

where:

rel_m is the graded relevance of the result at position m

Z_k is a normalization factor (perfect ranking at k = 1)

5.7.2 Automated Ranking Models

We employ different components of feature vectors, resulting in different ranking models. We adopt an intuitive and exploratory combination of features, with an aim to produce accurate building rankings.

Frequency-based Models

The normalized number of *favorites* for each building and the normalized number of *comments* for each building (accumulated from the metadata of Flickr images relevant to the building), independently form the basis of the *Frequency Models*. This means that according to the *Frequency Model*, each feature vector corresponding to a building consists of a single component; the normalized number of *favorites* or *comments*.

Polarity Model

Corresponding to each building, the *Polarity Model* utilizes feature vectors with two components; the positive and negative polarities as obtained from EmoLex.

Enhanced Sentic Model

In the *Enhanced Sentic model* the feature vectors comprise of 12 features. Apart from the 10 sentic feature components that are generated by using EmoLex, we also introduce the normalized number of *favorites* for each building and the normalized number of *comments*.

Filter Model

The *Filter Model* also comprises of 12 features. It uses the influence factors we determined earlier, and filters the data profiles $Profile(b_i)$, that we create for each building b_i.

As a first step, we build a *Bag of Words* (BoW) for each influential factor f_i in F, corresponding to the building type t_i. In order to do so, we exploit the Natural Language Toolkit (NLTK) WordNet package for Python [3, 23]. Using the WordNet package, we can derive the related words corresponding to each influential factor through the WordNet synsets. A synset or synonym ring is defined as a set of one or more synonyms that are interchangeable in some context without changing the truth value of the proposition in which they are embedded [23]. For example, for the Influential Factor, *Size of the building/structure*, we use the WordNet synsets to derive a BoW that are related to 'size'. We also use the Big Huge Thesaurus[25] API in order to extend the BoW.

The Big Huge Thesaurus is leveraged to extract synonyms, antonyms, related terms, similar terms and user suggestions in order to further extend the BoW. Finally, we manually go through the BoW for the different building types and further extend the BoW by including words that may be contextually relevant to the task at hand. For example, in case of the influential factor *Personal experiences*, we additionally include words that are not already in the BoW but might represent the context; like 'believe', 'feel', 'think' and so on.

In the second phase, we exploit the extended BoW, in order to filter the building data profiles $Profile(b_i)$, that we created for each building b_i of building type t_j. By doing so, we further prune the data by getting rid of potential noise. Figure 3 depicts this vital role played by the influential factors during the pre-processing stage.

Weighted Model

The *Weighted Model* is an extension of the *Filter Model*. Here, we consider the degrees of influence of each influential factor corresponding to the building type. First, we generate feature vectors using EmoLex for all buildings in the dataset, after pruning the building data profiles $Profile(b_i)$ corresponding to each influence factor. Then, the feature vectors are weighted with respect to their percentage of influence (depending on the building type), normalized and combined. The resulting weighted feature vectors are then used to train and test the model. In this way, the influence factors identified for each building type play a crucial role in the performance of the model itself.

As described earlier, we formulate the perception of a building as $Perception(b_i)$, and employ our ranking models to arrive at building rankings.

[25] http://words.bighugelabs.com/

Table 11: DBpedia properties that are used to materialize corresponding Influence Factors.

Influence Factors	Airports	Bridges	Churches	Halls	Skyscrapers
History Associated, Size, Materials Used, Level of Detail, Surroundings	dbpedia-owl: runwaySurface, dbpedia-owl: runwayLength, dbprop: cityServed, dbpedia-owl: locatedInArea, dbprop:direction [26]	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, dbpedia-owl: floorCount

5.8 Mining the Web to Correlate Influence Factors with Relevant Structured Data

Having overcome the first hurdles of 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 the influence factors with additional relevant information that can be extracted from DBpedia. Our approach to derive patterns in the perception of well-received structures is depicted in the Figure 6.

Table 12: Coverage of properties extracted from DBpedia for different architectural structures in our dataset.

Airports	Bridges	Churches	Halls	Skyscrapers
runwayLength: 95%	length: 67.79%	architectureStyle: 36.69%	seatingCapacity: 65.67%	floorCount: 91%

We exploit structured data from the DBpedia knowledge graph in order to correlate the influential factors with concrete properties and values. Table 11 depicts some of the properties that are extracted from the DBpedia knowledge graph in order to correlate the influence 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. In order to extract relevant data from DBpedia for each structure in our dataset, we first collect a pool of properties that correspond to each of the influence factors as per the building type (as shown in Table 11). In the next step,

[26] In Table 11 dbprop:direction, direction is one of north, south, east, west, northeast, northwest, southeast, or southwest

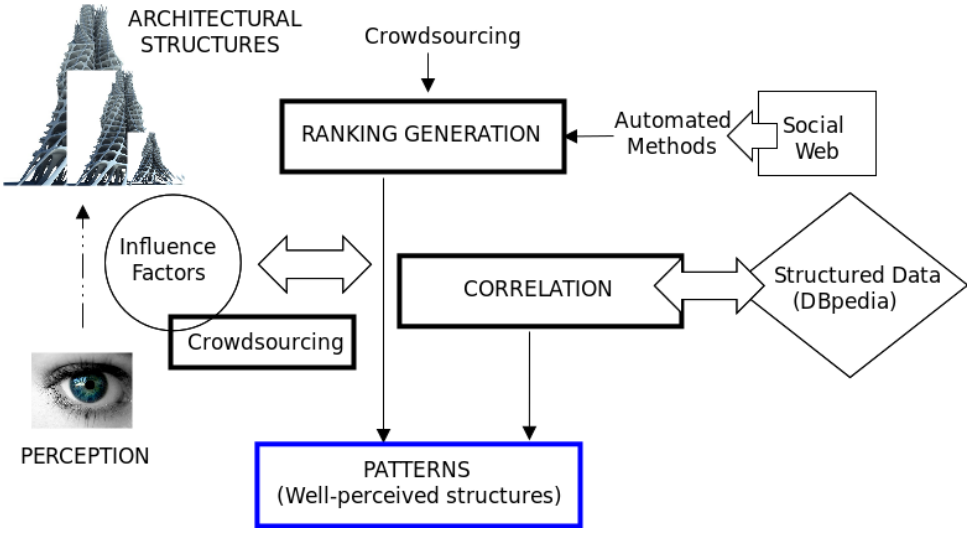


Figure 6: Approach to arrive at patterns for well-perceived architectural structures.

by traversing the DBpedia knowledge graph leading to each structure in our dataset, we try to extract corresponding values for each of the properties identified. The properties thus extracted semi-automatically, are limited to those available on DBpedia. In addition, it is important to note that not all structures of a particular type have the same properties available on DBpedia. Therefore, although all the identified values accurately correspond to the structure, the coverage itself is restricted to the data available on DBpedia (see Table 12).

5.9 Results & Evaluation

In this section, we present our dataset for experiments, results and evaluate the performance of our ranking models.

5.9.1 Dataset

As described earlier, we create building-type specific datasets and generate a new ground truth by exploiting crowdsourcing platforms like CrowdFlower[27], and LimeSurvey[28]. For our experiments, we consider the following architectural structure types : *Airports*,*Bridges*,

[27] http://crowdflower.com/
[28] http://www.limesurvey.org/en/

Churches, *Halls*, and *Skyscrapers*. We consider these building types since they are the most commonly found building types across different cities, as observed from Emporis[29], a real estate data mining company which is an authority on building data. The dataset we thereby created, consists of structures in the 10 biggest cities in Germany and USA (we choose USA and Germany due to the high social media traffic).
In order to ensure little variance in terms of the number of images per building, we only consider those buildings which correspond to at least a *threshold* number of images. Table 13 depicts the number of images, favorites and comments corresponding to each building type. We merge the textual metadata from the Flickr images (title, description and comments), for each image corresponding to each building, b_i. This constitutes the building profile $Profile(b_i)$ for each building.

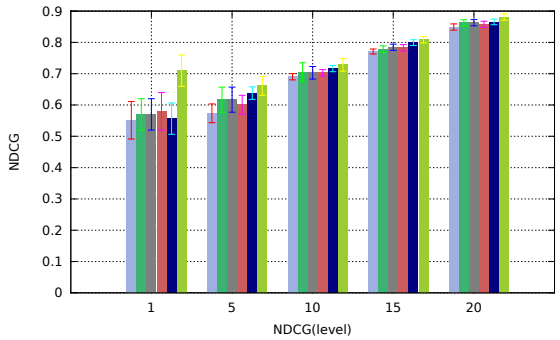
Table 13: Type-Specific Dataset Characteristics.

Building Type	# Buildings	# Images	# Favorites	# Comments
Airports	100	32,757	28,139	18,819
Bridges	59	12,050	19,281	25,677
Churches	139	28,683	20,857	37,036
Halls	67	20,178	11,676	14,271
Skyscrapers	178	61,538	138,899	183,051
Total	543	155,206	218,852	278,854

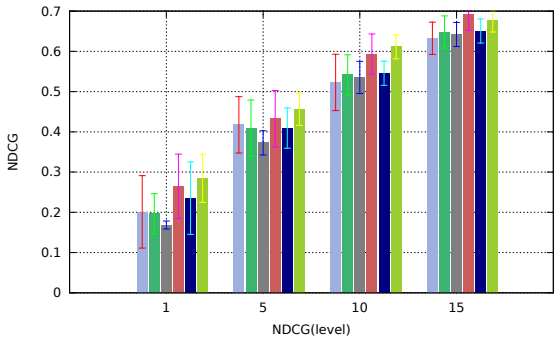
5.10 Performance of Ranking Models

Having established ground truths for the datasets pertaining to each of the building types, we evaluate our ranking models in order to observe their performance. The histograms in Figure 7 present the performance of our ranking models for the different building types. We plot NDCG values (averaged from 10 rounds of training and testing) at all levels.
We find that across the different building types, the *Weighted Model* outperforms other models at almost all NDCG levels (as illustrated in Table 14 for Halls). We infer that this performance gain is due to the weighted combination of feature vectors corresponding to a building, according to the influence factors. The cases bearing exceptions are discussed further below.
In Figure 7(a), we observe high NDCG values at all levels. This can be attributed to our observation that metadata from *airport* images on Flickr are highly rich with relevant emotion-contexts. The *Weighted Model* significantly outperforms the other models at

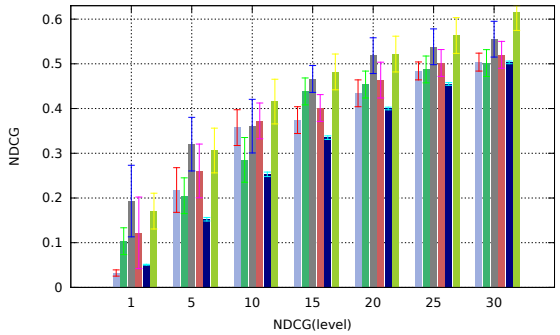
[29] http://www.emporis.com



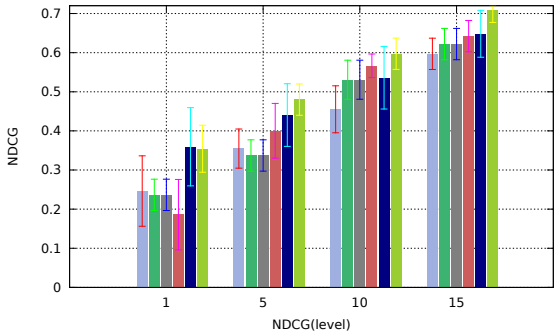
(a) Performance comparison of different ranking models for Airports.



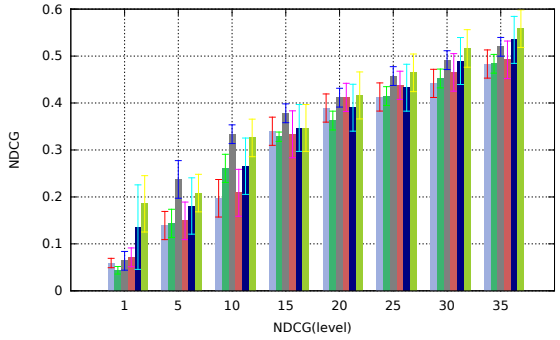
(b) Performance comparison of different ranking models for Bridges.



(c) Performance comparison of different ranking models for Churches.



(d) Performance comparison of different ranking models for Halls.



(e) Performance comparison of different ranking models for Skyscrapers.

Figure 7: Performance comparison of different ranking models for different building types.

NDCG@1. It marginally outperforms the other models at all the other levels of NDCG measured.

The Figure 7(b) presents the performance of our ranking models for the building type, *bridges*. In case of the *bridges*, we find that the *Enhanced Sentic Model* performs better than the other models at NDCG@15. The *Weighted Model* results in a better performance than the *Frequency-based Models* and the *Polarity Model*. The chart in Figure 7(c), depicts the case of *churches*, where the *Frequency-based Model* with *Avg. Comments* outperforms the other models at the NDCG levels 1 and 5. The *Weighted Model* performs the best at the remaining NDCG levels measured. In case of *halls*, as shown in Figure 7(d) we observe that the *Weighted Model*, followed by the Enhanced Sentic Model clearly outperform the *Frequency-based models* as well as the *Polarity Model*. Figure 7(e), presents the performance of the ranking models pertaining to *skyscrapers*.

An important revelation is that simple models based on reliable features like the normalized number of *favorites* and *comments* can perform fairly well. However, we need sophisticated models like the *Weighted Model* in order to attain a higher and more stable performance across different types of structures.

Table 14: Performance comparison of different ranking models for Halls.

Avg. NDCG@	Polarity Model	FM (Avg. Favorites)	FM (Avg. Comments)	Enhanced Sentic Model	Filter Model	Weighted Model
1	0.2462	0.2366	0.2366	0.1860	0.3595	0.3544
5	0.3547	0.3372	0.3372	0.4003	0.4405	0.4799
10	0.4552	0.5308	0.5308	0.5664	0.5359	0.5971
15	0.5970	0.6219	0.6219	0.6421	0.6482	0.7073

By using the *Standard Error* measure for statistical significance, we observe that the results are statistically significant at most NDCG levels. We observe a clear variance in the performance of different models across the different types of architectural structures. We attribute these differences to the varying importance of different emotions (which are used as features in training the models) with respect to different structure types. In addition, it is assumed that the architectural relevance of comments vary heavily among building types. For instance, while in case of churches, Flickr images and comments might likely be about the building itself, in case of bridges or airports, a large proportion of comments (and extracted sentiments) might indeed relate to other aspects. This leads to a more general finding about the need for filtering social media based on its relevance to the investigated buildings. While comments and extracted sentiments might relate to aspects independent of the depicted building (for instance, the photographic quality or an

event taking place at the depicted venue), additional pre-processing is required in order to better select social media of relevance for the task at hand.
The crowdsourced ground truths for different architectural structures and the detailed performance of our automated ranking models are additionally published for reference[30]. We publish our dataset abridged with the normalized popularity scores in the form of Linked Data by following the Linked Data principles. The knowledge base thus created, can be accessed and queried using our SPARQL endpoint [31].

5.11 Consolidation of Patterns: Proof-of-Concept

By correlating the influence factors to specific DBpedia properties, we can identify patterns for well-perceived architectural structures. In order to demonstrate how such observed patterns for architectural structures can be consolidated, we choose the influence factors, *Size* of the structure and *Level of Detail*. Although, this approach can be directly extended to other influence factors and across different kinds of architectural structures, in this first version of the Deliverable, we restrict ourselves to showcasing these influence factors.

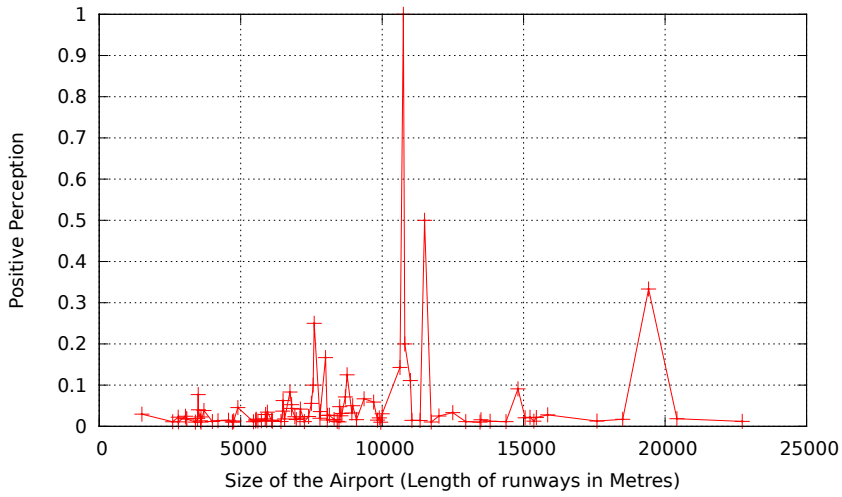
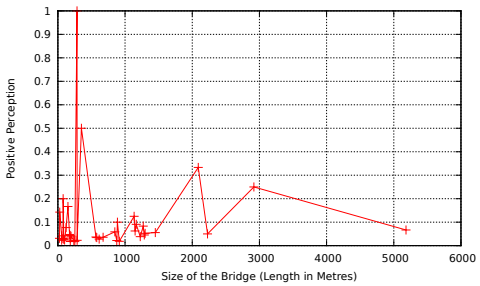


Figure 8: Influence of Size(total length of runways) in the perception of Airports.

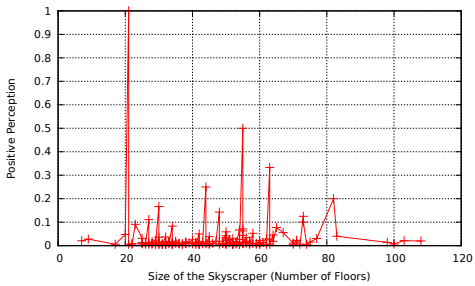
Airport 'size' is traditionally judged either by the number of operations (takeoffs and landings, runways) or the passenger traffic (number of passengers who fly in or out of the facility)[32]. Characteristics of major airports include two or more long runways capable of

[30] http://data-observatory.org/building-perception/
[31] http://meco.l3s.uni-hannover.de:8829/sparql
[32] http://virtualskies.arc.nasa.gov/airport_design/3.html

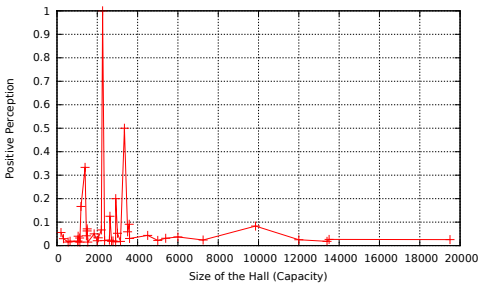
handling the larger jet airliners. The length of the runways are a fair indicator of the size of an airport. We observe that for each airport, we can extract indicators of size using the DBpedia property `dbpedia-owl:runwayLength`. We extract the length of the runways for each airport in our dataset in order to analyse and determine the well-received pattern for airports with respect to their size. The graph in Figure 8 shows how the popularity, i.e. the positive perception (as a factor of rank) of airports varies with their size. We observe that airports possessing runways with a length between 7,000-12,000 metres are generally well-perceived by people (higher $Perception(b_i)$).



(a) Influence of Size in the perception of Bridges.



(b) Influence of Size(# Floors) in the perception of Skyscrapers.



(c) Influence of Size in the perception of Halls.

Figure 9: Influence Factors (size) in the perception of Architectural Structures.

Similarly, in case of bridges the influence factor 'size' can be represented using the DBpedia properties `dbpedia-owl:length`, `dbpedia-owl:width` and `dbpedia-owl:mainspan`, for halls we can use the DBpedia properties `dbprop:area` and `dbprop:seatingCapacity`, while we can use `dbpedia-owl:floorCount`, and `dbprop:height` to consolidate the well-perceived patterns for Skyscrapers. We thereby extract corresponding property values for each structure in our dataset using the DBpedia knowledge graph.

Figure 9(a) shows how the popularity, i.e. the positive perception of bridges varies with their size (in terms of length of the bridge). It is interesting to note that long bridges are not necessarily perceived well. We note that some bridges with length less than 1000 metres are perceived very well by people ($Perception(b_i)$>0.5). The graphs in Figure 9(b) shows that skyscrapers having 25-65 floors form the crux of the most well-perceived skyscrapers. We observe that halls with a seating capacity between 1000-4000 people are well-perceived with the positive perception varying between 0.1 and 1.

For churches, we demonstrate the consolidation of patterns with respect to the influence factor *Level of Detail*. The `dbprop:architectureStyle` is a good measure of the detail in the structure. We thereby correlate the influence factor *Level of Detail* with the architecture styles using `dbprop:architectureStyle` in the DBpedia graph. By doing so, the churches in our dataset are mapped to 15 different architectural styles. The 3 most popular styles are found to be 'Gothic Revival', 'Romanesque', and 'Gothic'.

We have shown how architectural patterns can be mined by correlating structure features with properties from DBpedia. It is very important to note that the architectural patterns observed and presented here are based on merely a single dimension (i.e., size or level of detail). We have already showed that perception of an architectural structure involves multiple factors. In order to establish more concrete, meaningful and thorough architectural patterns, we will consider the remaining influential factors in a similar manner for each type of structure.

5.12 Caveats and Limitations

Apart from the previously mentioned challenge of ensuring appropriate context of the metadata (especially *comments*) from Flickr images, we bring to light the other aspects in our work that can be upheld as possible caveats.

The images of structures displayed to crowd workers during the process of building ground truths and assessing influence factors, predominantly depict exterior views (while functionality of certain structures may be more dependent on the interior settings of the structures). Our usage of exterior views was driven by the intuition that architectural structures are generally more recognizable from their external rather than internal appearance. In addition to this, providing an exterior view of the structure with a representational excerpt of the immediate surroundings would aid the crowd workers.

The architectural patterns that can be extracted here are reliant on the relevant properties

available on DBpedia corresponding to each structure. This means that the patterns thus mined are limited to the knowledge available from the source.

5.13 Conclusions & Future Work

One of our main contributions is the pipeline we designed that can be tailored to specific architectural structure types in order to allow the measurement of public perception of structures. Alongside this, the influence factors and the ground truths established for different types of architectural structures are key contributions of our work. By exposing the data we generated as Linked Data, we make it available for public use. An interactive visualization supports further deliberation[33].

Through our experiments, we find that in the task of ranking structures based on their associated perception extracted from Web data, a big challenge is to ensure the relevance of the extracted text to the structure-type that we are interested in. We are led to believe that pruning relevant data to closely fit the corresponding structure types will have a positive impact on ranking performance. In this respect, filtering mechanisms which consider the most fine-grained type possible (for instance, airport instead of building), seem the most promising. This is due to the insight that different types are usually influenced by different factors, as identified through our crowdsourcing activities. To this end, influence factors can provide a means to tailoring NLP-based filtering methods.

A broad range of architectural insights can be facilitated as a result of rankings thus generated. We demonstrate this by correlating with building characteristics extracted from DBpedia. Our models and methods can help in analysing the evolution of the popularity of a building. Apart from architects, builders, magazines, News Channels, building corporations or other parties interested in building rankings, can greatly benefit from this approach; by eliminating a large amount of human costs, otherwise required to arrive at such rankings. In addition, our approach to crowd-source the influential factors further reduces the manual labour and need for cumbersome human intervention. In many cases, influential factors with respect to different structure types are not known apriori.

In the imminent future, one direction for investigations is the correlation of building with additional structured data. With respect to mining architectural patterns, we will extend our work to cover a rigorous analysis that can help us mine patterns with multiple facets. For example, to mine patterns like 'skyscrapers with x size, y uniqueness, and z materials

[33] https://a.tiles.mapbox.com/v3/ujwal07.4qu84cxr/page.html?secure=1#2/0/0

used are best perceived'. Since not all the architectural structures in our dataset have data for the associated properties in the DBpedia knowledge graph, we will further mine the Web in order to populate our knowledge base. In this way, we can extract concrete patterns with respect to different kinds of architectural structures, while encompassing all the related influence factors. Data from the Social Web can also be put to vital use, for example, using tweets from Twitter to manifest concrete statistics relevant to influence factors like *Personal Experiences* involving a structure or *Uniqueness* of a structure.

6 Decisions & Risks

6.1 Technical decisions and impacts

Web-based user interface for the Manual Interlink prototype

The graphical user interface of the interlinking software prototype ("InterlinkUI") is developed with a web technology stack running in a web browser. The browser environment implies advantages over a standalone desktop application, the most important one being the platform independence of the application. A web browser also provides a standardized environment[34]. developers can work with. This environment is (to the most degree) the same on different platforms, e.g. Microsoft Windows, Linux and MacOS, but also for the very popular mobile platforms Android, iOS, Windows Mobile, etc., which are running on mobile phones and tablets. This has the tremendous advantage that when developing an application with a web technology stack it will automatically run on the most popular desktop and mobile platforms, without the need to change the application code.

SPARQL End-point interface to external data sets for Manual Interlink prototype

For the Manual Interlink prototype it has been decided to base the navigation and interaction with data sets with external SPARQL interfaces. This not only keeps the implementation effort low compared to e.g. reading complete datasets into memory. It also makes the tool generic and adaptable to future vocabularies and out-sources scalability aspects to the vocabulary and data set providers themselves.
Event though the SPARQL standards 1.0 http://www.w3.org/TR/rdf-sparql-query/ and 1.1. http://www.w3.org/TR/sparql11-overview/ by the W3C organization have been widely accepted, their implementation specifics vary between end-point implementations and configurations, which might lead to ad-hoc circumventions on the prototype side.

[34]Client-side web standards are is organized in multiple standard bodies and working groups. The most prominent ones are the World Wide Web Consortium (W3C, http://w3.org/ and the Web Hypertext Application Technology Working Group (WHATWG, http://www.whatwg.org/

6.2 Risk Assessment

Risk Description The use of SPARQL endpoints is replaced by other standards and future versions of Linked Data are presented differently

Risk Assessment .

Impact High

Probability Low

Description Even though they differ in implementation details, SPARQL endpoints will very likely remain to play a role in the future of linked data. Additional layers such as security etc. might be added on top which would require adaptions of the prototypical tools described here.

Contingency Solution The organisations of the DURAARK consortium are closely following the developments of the Semantic Web and Linked Data communities. If severe modifications of elemental building blocks such as SPARQL endpoints are being introduced into the overall LD approaches, conceptual and technical migration paths will very likely be developed along side in many other research initiatives and products.

7 Licenses

The following table gives an overview of the software licences generated and used for the web services and UI modules implementation:

IPR Type	**IP used or generated**	**Software name**	**License**	**Information**
software	generated	Manual Interlinking prototype	MIT	D3.4
software	used	dagre - Graph layout for JavaScript	MIT	https://github.com/ cpettitt/dagre
software	used	D3.js	BSD	http://d3js.org/
software	used	Require.js	BSD or MIT	http://requirejs.org/
software	used	LDSpider	GNU Lesser GPL	https://code.google.com/p/ldspider

8 Conclusion and Impact

The prototypical software tools for Semantic Digital Interlinking and Clustering presented in the first part of this deliverable play a crucial role in the preservation processes of semantically rich building models: They allow the discovery, use and maintenance of external datasets and vocabularies to enrich the building models themselves as well as their metadata records in Digital Longterm Preservation systems such as DURAARK. Since both, the semantic description of buildings and their models themselves as well as the semantically rigid description of archival content is only slowly beginning to gain traction, the tools presented here help on very fundamental fundamental levels:

- The **Clustering Prototype** allows to discover relevant information regarding building and construction in existing and future vocabularies and datasets such as bSDD, DBPedia and Getty AAT. It allows targeted searches and crawls of linked data that will greatly enhance the on-going process of building up common and shared knowledge bases for the description of building and the preservation of related data.

- The **Manual Interlinking Prototype** allows to create, validate and asses semi-automatically or manually created links between resources compiled by the clustering mechanisms. With the tools presented here domain experts such as librarians, historians, architects, engineers and other building experts are enabled to related the heterogeneous and disparate knowledge and dataresources existing today in a user-friendly manner.

- The **Mashing Prototype** and case study examples provided in this deliverable show examples of how such clusered and interlinked semantically rich datasets can be used to gain new insights e.g. into the perception of buildings that would be difficult to acquire by traditional means. The extensive descriptions of the methods and tools can be adapted to other search and analysis scenarios that could e.g. help architects, planners and other stakeholders to gain more insight into the impact of design decisions.

This D3.4 deliverable milestone will be followed by extensive case studies and validations of clustering and interlinking by practitioners during workshops and on crowd sourcing platforms. The stand-alone tools presented here will be seamlessly embedded into the

DURAARK workbench framework and UI components in later milestone releases but are useful in their on right in other Linked Data scenarios.

References

[1] Christopher Alexander. *The timeless way of building*, volume 1. Oxford University Press, 1979.

[2] Christopher Alexander, S Ishikawa, and M Silverstein. Pattern languages. *Center for Environmental Structure*, 2, 1977.

[3] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[4] Philomena M Bluyssen, Christian Cox, Nadia Boschi, Marco Maroni, Gary Raw, CA Roulet, and Flavio Foradini. European project hope (health optimisation protocol for energy-efficient buildings). In *Healthy Buildings*, volume 1, pages 76–81, 2003.

[5] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*, 2011.

[6] R. J. Brachman. What IS-a is and isn't: An analysis of taxonomic links in semantic networks. *Computer*, 16(10):30–36, 1983.

[7] Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. *Journal of Machine Learning Research-Proceedings Track*, 14:1–24, 2011.

[8] Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P Sheth. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *ICWSM*, 2012.

[9] WK Chow. Proposed fire safety ranking system eb-fsrs for existing high-rise nonresidential buildings in hong kong. *Journal of architectural engineering*, 8(4):116–124, 2002.

[10] Wei-Ta Chu, Yu-Kuang Chen, and Kuan-Ta Chen. Size does matter: how image size affects aesthetic perception? In *Proceedings of the 21st ACM international conference on Multimedia*, pages 53–62. ACM, 2013.

[11] Sanjiv R Das and Mike Y Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

[12] Victor de Boer, Michiel Hildebrand, Lora Aroyo, Pieter De Leenheer, Chris Dijkshoorn, Binyam Tesfa, and Guus Schreiber. Nichesourcing: Harnessing the power of crowds of experts. In *Knowledge Engineering and Knowledge Management*, pages 16–20. Springer, 2012.

[13] Harry Halpin, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson. When owl: sameas isn't the same: An analysis of identity in linked data. In *The Semantic Web–ISWC 2010*, pages 305–320. Springer, 2010.

[14] Bill Hillier. Space is the machine: a configurational theory of architecture. 2007.

[15] Robert Isele, Jürgen Umbrich, Chris Bizer, and Andreas Harth. LDSpider: An open-source crawling framework for the web of linked data. In *Proceedings of 9th International Semantic Web Conference (ISWC 2010) Posters and Demos*, 2010.

[16] Glen Jeh and Jennifer Widom. Simrank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. ACM.

[17] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226, New York, NY, USA, 2006. ACM.

[18] Lyndon Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proceedings of the 15th international conference on Multimedia*, pages 631–640. ACM, 2007.

[19] Margrethe Kobes, Ira Helsloot, Bauke de Vries, and Jos G Post. Building safety and human behaviour in fire: A literature review. *Fire Safety Journal*, 45(1):1–11, 2010.

[20] Maria Kordjamshidi. *House Rating Schemes*. Springer, 2011.

[21] Neal Lathia, Daniele Quercia, and Jon Crowcroft. The hidden image of the city: sensing community well-being from urban mobility. In *Pervasive Computing*, pages 91–98. Springer, 2012.

[22] Kevin M Leyden, Abraham Goldberg, and Philip Michelbach. Understanding the pursuit of happiness in ten major cities. *Urban Affairs Review*, 47(6):861–888, 2011.

[23] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[24] Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34. Association for Computational Linguistics, 2010.

[25] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 2012.

[26] Edward Ng. *Designing high-density cities for social and environmental sustainability*. Earthscan, 2010.

[27] Bernardo Pereira Nunes, Stefan Dietze, Marco Antonio Casanova, Ricardo Kawase, Besnik Fetahu, and Wolfgang Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking. In *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, pages 548–562, 2013.

[28] Johan Oomen and Lora Aroyo. Crowdsourcing in the cultural heritage domain: opportunities and challenges. In *Proceedings of the 5th International Conference on Communities and Technologies*, pages 138–149. ACM, 2011.

[29] Ian Oppenheim, Heiner Mühlmann, Gerhard Blechinger, Ian W Mothersill, Peter Hilfiker, Hennric Jokeit, Martin Kurthen, Günter Krämer, and Thomas Grunwald. Brain electrical responses to high-and low-ranking buildings. *Clinical EEG and Neuroscience*, 40(3):157–161, 2009.

[30] Ilan Oppenheim, Manila Vannucci, Heiner Mühlmann, Rainer Gabriel, Hennric Jokeit, Martin Kurthen, Günter Krämer, and Thomas Grunwald. Hippocampal contributions to the processing of architectural ranking. *NeuroImage*, 50(2):742–752, 2010.

[31] Claudia Orellana-Rodriguez, Ernesto Diaz-Aviles, and Wolfgang Nejdl. Mining emotions in short films: user comments or crowdsourcing? In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 69–70. International World Wide Web Conferences Steering Committee, 2013.

[32] Luis Perez-Lombard, Jose Ortiz, and Christine Pout. A review on buildings energy consumption information. *Energy and buildings*, 40(3):394–398, 2008.

[33] Daniele Quercia. Urban: crowdsourcing for the good of london. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 591–592. International World Wide Web Conferences Steering Committee, 2013.

[34] Daniele Quercia, João Paulo Pesce, Virgilio Almeida, and Jon Crowcroft. Psychological maps 2.0: A web engagement enterprise starting in london. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1065–1076. International World Wide Web Conferences Steering Committee, 2013.

[35] C-A Roulet, F Flourentzou, HH Labben, M Santamouris, I Koronaki, E Dascalaki, and V Richalet. Orme: A multicriteria rating methodology for buildings. *Building and Environment*, 37(6):579–586, 2002.

[36] Johner N. Flourentzous F. Greuter G. Roulet, C.-A. European project hope (health optimisation protocol for energy-efficient buildings). In *Healthy Buildings*, volume 1, 2003.

[37] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, pages 891–900. ACM, 2010.

[38] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. Analyzing and predicting sentiment of images on the social web. In *Proceedings of the international conference on Multimedia*, pages 715–718. ACM, 2010.

[39] Camillo Sitte. *City planning according to artistic principles*. Rizzoli, 1986.

[40] Louis H Sullivan. *The autobiography of an idea*, volume 281. Courier Dover Publications, 1956.

[41] Kwok Tung Tsang. Stochastic quantitative fire risk assessment on old buildings in hong kong. 2012.

[42] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[43] Yulan Yang, Baizhan Li, and Runming Yao. A method of identifying and weighting indicators of energy efficiency assessment in chinese residential buildings. *Energy Policy*, 38(12):7687–7697, 2010.

[44] Hugo Zaragoza, Henning Rode, Peter Mika, Jordi Atserias, Massimiliano Ciaramita, and Giuseppe Attardi. Ranking very many typed entities on wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 1015–1018. ACM, 2007.